

---

**Using differential MNase-seq to dissect nucleosome  
organization around promoters and enhancers in  
*D.melanogaster***

Alessio Renna

---

München 2018



Dissertation zur Erlangung des Doktorgrades  
der Fakultät für Chemie und Pharmazie  
der Ludwig-Maximilians-Universität München

**Using differential MNase-seq to dissect nucleosome  
organization around promoters and enhancers in *Drosophila  
melanogaster***

Alessio Renna

aus

Taranto, Italien

2018





## **Erklärung:**

Diese Dissertation wurde im Sinne von §7 der Promotionsordnung vom 28. November 2011 von Herrn Prof. Dr. Karl-Peter Hopfner betreut.

## **Eidesstattliche Versicherung:**

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, den 17/07/2018

---

Alessio Renna

Dissertation eingereicht am 19/04/2018

1. Gutachter: Prof. Dr. Karl-Peter Hopfner

2. Gutachter: PD Dr. Philipp Korber

Mündliche Prüfung am 28/06/2018



# Table of Contents

<b>I</b>	<b>ACKNOWLEDGMENTS .....</b>	<b>IV</b>
<b>II</b>	<b>ABSTRACT .....</b>	<b>VI</b>
<b>1</b>	<b>INTRODUCTION.....</b>	<b>1</b>
1.1	Gene regulation and chromatin .....	1
1.2	Nucleosome definitions.....	3
1.3	Determinants of nucleosome mapping.....	5
1.4	DNA sequence and nucleosome mapping .....	5
1.4.1	Di-nucleotide 10 bp periodicity .....	6
1.4.2	Homopolymeric DNA sequences.....	7
1.4.3	G + C content .....	7
1.5	Nucleosome motions and spontaneous unwrapping .....	8
1.6	Trans-factors and nucleosome mapping .....	10
1.7	The core promoter and its shape .....	12
1.8	Pol II pausing and nucleosome barrier.....	15
1.9	Probing genome accessibility.....	16
1.9.1	DNase-seq and enhancers .....	16
1.9.2	MNase-seq and nucleosome mapping.....	17
1.10	MNase-seq biases.....	20
1.11	Differential MNase-seq (Fragile/Resistant nucleosomes) .....	23
<b>2</b>	<b>AIM OF THE THESIS .....</b>	<b>27</b>
<b>3</b>	<b>METHODS .....</b>	<b>29</b>
3.1	Cell biology, molecular biology and biochemical procedures .....	29
3.2	High-throughput genome-wide procedures.....	31
3.3	Computational procedures .....	37
<b>4</b>	<b>RESULTS.....</b>	<b>40</b>
<b>4.1</b>	<b>Study design.....</b>	<b>40</b>
4.1.1	Experimental design.....	40
4.1.2	CREs definitions .....	43

## Table of Content

<b>4.2</b>	<b>PART I – MNase biases</b>	<b>44</b>
4.2.1	MNase sequence bias at the cut site is complex	44
4.2.2	MNase sequence bias in chromatin is reduced	45
4.2.3	MNase sequence bias does not determine nucleosome mapping around promoters	48
4.2.4	MNase digestion on FA cross-linked chromatin	51
4.2.5	MNase digestion within the nucleosome is mostly discrete	53
4.2.6	MNase digestion within the nucleosome is mostly asymmetric	57
4.2.1	The +1 nucleosome of BP genes is asymmetrically digested by MNase	62
<b>4.3</b>	<b>PART II – Differential MNase-seq</b>	<b>65</b>
4.3.1	Dinucl- and sub-nucleosomes provide new information on nucleosome landscape	66
4.3.2	Differential MNase-seq can distinguish nucleosome populations	70
4.3.3	Nucleosome populations can be identified within one digestion level	73
4.3.4	Nucleosome populations have distinct DNA features	74
4.3.5	The one-digestion level approach is quite robust against digestion level	77
4.3.6	Nucleosome populations are located differently around CREs	77
4.3.7	Alternative scores for nucleosome population distinction: mono- and oligo- based scores (monos and oligos)	82
4.3.8	Mono- and oligo- scores better capture local chromatin accessibility	83
4.3.9	Genes closer to each other show low expression plasticity in <i>D.melanogaster</i> and vice versa	86
4.3.10	Genes with high and low expression plasticity show different GC profile around promoters	89
4.3.11	Sequence and chromatin landscapes are connected in divergent promoters	92
<b>4.4</b>	<b>PART III – MNase-Short-ChIP-seq</b>	<b>95</b>
4.4.1	MNase-Short-ChIP-seq: a new approach to better measure chromatin features in active genomic regions	95
4.4.2	Histone mark occupancy is influenced by activity and distance of closer genes	100
<b>5</b>	<b>DISCUSSION</b>	<b>103</b>
5.1	MNase biases	103
5.1.1	MNase sequence bias does not determine nucleosome mapping	103
5.1.2	Evidence of MNase digestion within the nucleosome driven by nucleosome unwrapping	104
5.1.3	The +1 nucleosome is asymmetrically digested by MNase	105
5.2	Differential MNase-seq	106

## *Table of Content*

5.2.1	Differential MNase-seq: a powerful tool to study the nucleosome organization.....	106
5.2.2	Differential MNase-seq: a new one-digestion level method.....	107
5.2.3	Functional plasticity correlates with a dual nucleosome organization.....	109
5.2.4	Divergent promoters: a special gene organization .....	111
5.2.5	Gene expression plasticity is mirrored in the underlying GC content .....	112
5.3	MNase-Short-ChIP-seq.....	113
<b>6</b>	<b>OUTLOOK.....</b>	<b>115</b>
<b>7</b>	<b>APPENDIX.....</b>	<b>118</b>
7.1	Supplementary Figures .....	118
7.2	Abbreviations .....	129
<b>8</b>	<b>BIBLIOGRAPHY .....</b>	<b>131</b>

# I ACKNOWLEDGMENTS

I start my acknowledgments by mentioning my supervisor Ulrike Gaul. She gave me the possibility to join her group, the freedom to express my creativity in the lab and the opportunity to further develop as a scientist. I am really grateful to her for guiding me during these years.

I also thank Prof. Karl-Peter Hopfner for being my thesis supervisor. His availability was vital to accomplish the graduation. In this regard, I am also thankful to the other TAC members: Prof. Jürg Müller, Prof. Philipp Korber and Dr. Stefan Canzar. Their suggestions and critical comments on my work were really appreciated and helpful to refine this project. Finally, I thank for their time the board examination members that I have not mentioned so far: Prof. Julian Stinglele, Prof. Veit Hornung and Prof. Roland Beckmann.

This MNase project, as we called in the lab, was quite a journey. It started as a control experiment then it developed further by following several interesting findings. For that reason, here, I have to thank Mark Heron and Ulrich Unnerstall. Their support for this project was essential by providing meaningful analysis and scientific discussions. In particular, I have to mention Mark since the continuous email exchange, about data and observations, was crucial to develop new ideas and to plan further analyses and experiments. Definitely, our collaboration was one of the best parts of my PhD. I am also thankful to Andrea “Ennio” Storti, Marta “Teresa” Bozek and Roberto Cortini for helping through DHS data. Naturally, I want to thank all the other Gaul lab members for creating a really nice working atmosphere. Our attitude in being supportive between each other was fruitful in exchanging ideas and lab experiences. In particular, I want to mention Miro Nikolov and Christine Rottenberger with whom I shared another project and many many hours in the cell culture. May the clones be with you, guys. Finally, a honorable mention goes to Sara Batelli, Stefano Ceolin, Zhan Qi (beyond the people already cited) for assisting my coffee addiction and for sharing little and juicy chats during lunch breaks.

I would like to express my gratitude to the IRTG SFB1064 Graduate School “Chromatin Dynamics”. It is a great school in which I got to know amazing people and friends that enriched my PhD. A particular thanks goes to the IRTG coordinator Dr. Elizabeth Schroeder-Reiter for her professionalism and kindness.

Usually, job-related acknowledgments are followed by mentioning people important in private life. I consider myself very lucky that I could copy this part from my master thesis. They are still the foundations on which I am building my life, with two significant add-ons. Starting with my family, what can I say about them. They are great, they love me unconditionally and I felt it during all the PhD years. Many people describe the PhD as an emotional roller coaster with many ups and downs. In that regard, in both situations, my family was a terrific moral support with their optimism. It is something I learned not to take it for granted. Especially, I want to thank once again my sister Patrizia that pushed me, with her phone calls and visits, to further open my mind to other things that are not lab-related. The 2018 is a special year for her, consequently, for me too. Indeed, I am super happy that I can be called uncle now. I hope that I will be a good one for Jiji.

During the PhD, I have learned more and something new about myself. Recently, this path led me to Veronica. I want to thank her just for what she is. We love and simply understand each other and I am very lucky to find her by my side every day. I thank Veronica for the patience and understanding demonstrated during the thesis writing period, which can be quite stressful. More in general, I want to thank her for the immense and daily support I received in these last two years.

Life is more colorful with friends. Therefore, I have to mention the pezzavillians in Taranto. We share a lot of memories together since childhood. I hope we will have many other possibilities to create new ones. Especially, I send my greetings to Carlo, Titti, Mary and Daniele and I want to thank them for their closeness, even if we are separated by many kilometers. Finally, I would like to mention Andrea again (and with this we are even) for sharing so many moments in the lab and outside of it that I cannot count. We sincerely supported each other over these years.

At the end, I would like to express my sincere gratitude to myself. As I wrote for my master thesis, I will pat my shoulder, and, with a grin, I will say “well done, Alessio”.

## II ABSTRACT

Regulatory proteins compete with nucleosomes for access to the DNA. Therefore, nucleosome organization around promoters and enhancers strongly influence how these *cis*-regulatory elements (CREs) function. Currently, MNase-seq is still the method of choice for nucleosome mapping. However, it is affected by three biases: 1) preference of cutting at the A or T; 2) ability to cut within the nucleosome; 3) dependency on the degree of the digestion. In particular, nucleosomes show a differential sensitivity to MNase digestion that defines distinct nucleosome populations. MNase-sensitive or fragile nucleosomes are strongly susceptible to MNase digestion, while insensitive or resistant nucleosomes are less susceptible to digestion. Thus, differential MNase-seq, in which nucleosome organization is probed by varying digestion conditions, is a powerful tool to assess chromatin properties.

The present study applies a novel differential MNase-seq approach in *D.melanogaster* S2 cells, by analyzing the information contained in oligo-, mono- and sub-nucleosome fractions. This information is used to better characterize MNase-seq biases, to develop a new method for the definition of nucleosome fragility and resistance, and to better measure chromatin related features on active genomic regions (MNase-Short-ChIP-seq).

Firstly, the study shows the complexity of the sequence content around MNase cut sites and it rejects the hypothesis that MNase sequence bias determines how nucleosomes are mapped around promoters. Furthermore, our results indicate that MNase digestion within the nucleosome is mostly driven by the biophysical properties of the DNA-histone interactions. Indeed, intra-nucleosome digestion, found to be mostly discrete and asymmetric, strongly correlates with DNA-histone binding properties as assessed by published biophysical studies.

Secondly, we developed a new differential MNase-seq method to distinguish nucleosome populations based on the mono- to sub-nucleosome ratio within a single digestion level. Thus, our method focuses on the MNase sensitivity of individual nucleosomes by capturing the DNA-histone binding strength. Moreover, the mapping of nucleosome populations defined in this fashion reveals novel features of the nucleosome organization around CREs. In particular, the dualism between CREs with high and low functional plasticity is associated with differences in fragility and resistance landscapes, in which sequence and activity components interplay.



## *Abstract*

Finally, we propose MNase-Short-ChIP-seq, which captures MNase-sensitive nucleosomes better than the standard MNase-ChIP approach and strongly improves the assessment of chromatin-related features on active genomic regions. Moreover, thanks to this method, our study finds that the deposition of histone tail PTMs around promoters is influenced by the activity and distance of neighboring genes.

Overall, the present study represents the most comprehensive investigation of MNase digestion and, by using a novel differential MNase-seq approach, the most detailed characterization of the *D.melanogaster* chromatin landscape around CREs available to date.



# 1 INTRODUCTION

## 1.1 Gene regulation and chromatin

Control of gene expression is essential for all living organisms and it is involved in all physiological processes, such as cell growth, development, differentiation and maintenance. Gene expression regulation is a strict process which ensures that the genetic information required for a given biological function is expressed at appropriate levels in a correct space/time frame. This control is executed by DNA binding proteins called transcription factors (TFs) that recognize specific DNA sequences, known as DNA motifs, within genomic regions called as cis-regulatory elements (CREs). Therefore, CREs work as landing platform for the TFs. The overall function of this conserved and universal principle is to recruit the transcription machinery, at the transcription start sites (TSSs) of genes in order to initiate their expression.

The spatiotemporal control of gene expression is mainly determined by the action of TFs on distal CREs, such as enhancers. These interactions are mostly driven by upstream signaling pathways which in turn are activated by internal, such as growth and development, and environmental stimuli. In a nutshell, enhancers define where the transcription machinery should be located along the genome so which genes should be activated, priming the assembly of the transcriptional machinery. The core promoter is essential for regulating the gene expression level. They are CREs of around 150 base pair (bp) that directly surround the TSS. Each gene can have one or few core promoters (called alternative promoters). The combination of DNA motifs and DNA biophysical properties constitute the architecture of core promoters. Core promoter architectures determine the rate of recruitment of the transcription machinery, and therefore the level of gene expression plasticity.

The genome of an organism, between  $1.70 - 1.93 \times 10^8$  bp in *Drosophila melanogaster* (Ellis et al., 2014), has to fit in nuclei of few  $\mu\text{m}$ . The DNA packaging inside the nuclei is assisted by small basic proteins named histones, assembling in a complex structure named chromatin. The basic repeating unit of the chromatin is the nucleosome. It is composed of a nucleosome core and a linker DNA. The nucleosome core consists of around 146-147 bp of DNA, called nucleosomal

DNA, wrapped around an histone octamer which is composed of two copies of each four core histone proteins: H2A, H2B (organized in heterotypical dimers), H3 and H4 (organized in a tetramer complex) (Luger et al., 1997). The interactions between nucleosomal DNA and histones are driven by direct charge-charge interactions, between the negative phosphate backbone of the DNA and the positive charged residues of the histones, hydrogen-bonds and non-polar interactions (Davey et al., 2002). Depending on the species and cell type up to 90% of the genome is organized in nucleosomes.

Between 25-30% of the histone mass is composed of relatively unstructured long tail domains that exit the nucleosome core. These long tail domains are located at the N-terminal of all four core histones and at the C-terminus of the histone H2A. The histone tails marginally contribute to the nucleosome core stability and are instead mostly involved in inter-nucleosome interactions necessary for the generation of higher order structures of the chromatin (Gordon et al., 2005). Nevertheless, a recent study showed that tailless nucleosomes, analyzed by crystal structures of nucleosomes assembled with tailless H3 or H2B histones, affect the strength of the DNA–histone interactions resulting in decreased nucleosome stability (Iwasaki et al., 2013). However, histones can act as substrates for numerous epigenetic post-translational modifications (PTMs) both in their core and tail domains. Histone tail PTMs are characterized by high abundance and a certain level of combinatorial complexity along the genome. Indeed, histone tail residues can easily act as substrates of many enzymes due to the great accessibility. These modifications affect chromatin condensation by direct and indirect mechanisms.

Nucleosomes are spaced by linker DNA of varying length that changes depending on species and cell type. The primary order structure of the chromatin is a simple concatenation of nucleosomes and is called 10-nm-fiber or “beads-on-a-string”, as visualized by electron microscopy. In some species the linker histone H1 or its variants are necessary for a further packaging of the chromatin. H1 binds the linker DNA in close proximity of the nucleosomal DNA. It stabilizes the nucleosome structure and together with other non-histone proteins facilitates the formation of chromatin fibers and their folding in higher order structures, such as 30nm fibers and chromosomes (Robinson and Rhodes, 2006). The structure composed of a nucleosome with one bound linker histone is called chromatosome and permits in most species a protection of around 160 bp from nucleases. Not all genomic regions are packed in identical manner, therefore open and closed domains can be distinguished depending on the level of

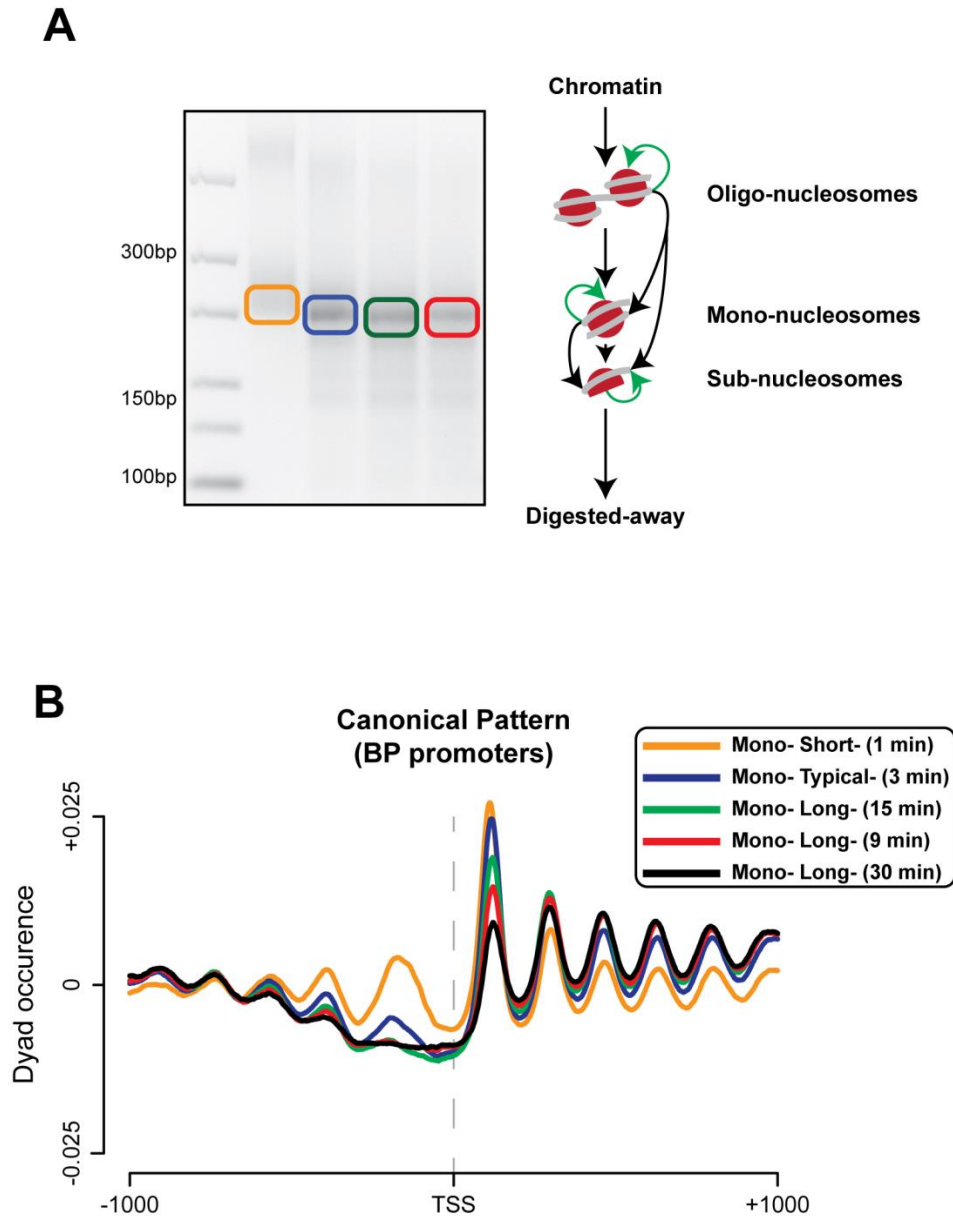
chromatin compaction. Packaging of the DNA in chromatin creates an evolutionary possibility for genes regulation. Indeed, a CRE within a nucleosome will be less accessible, reducing the probability of transcription of genes that it regulates. Learning how nucleosomes are organized around these elements and what are the mechanisms to overcome the nucleosome barrier has been driving a huge amount of studies in the last decades, with still a lot remains to be understood.

## **1.2 Nucleosome definitions**

I would like to spend here few words about nucleosome positioning, nucleosome occupancy and nucleosome canonical pattern definitions, due to their frequent occurrence in this dissertation (Kaplan et al., 2010a). Nucleosome positioning describes the location of a nucleosome along the genome. In a cell population in which the same genomic region has several possible nucleosome configurations (due to thermodynamics fluctuations and active ATP-consuming mechanisms), nucleosome positioning measures how often a bp coincides with a specific nucleosomal DNA coordinate. This coordinate can be the nucleosomal DNA entry (lowest coordinate), the exit (147<sup>th</sup> bp) or more frequently the center (73-74<sup>th</sup> bp). Nucleosome position is extremely relevant to study CREs activity, since slight variations of few bp can lead a DNA motifs to be within or closer to the entry/exit of the nucleosome, therefore changing its accessibility (Albert et al., 2007). Moreover, two levels of nucleosome positioning can be distinguished: translational and rotational positioning. The former describes the longitudinal location along the genome, whereas the latter indicates a cluster of alternative positions with a span of around 10 bp (one helical turn) upstream or downstream. On the other hand, nucleosome occupancy represents how frequently a base pair is covered by a nucleosome.

The canonical nucleosome pattern around TSSs is a nucleosome organization typically found in housekeeping genes. It is characterized by a nucleosome depleted region (NDR) of around 150 bp that includes the TSS. The NDR is surrounded by two well positioned nucleosomes: the -1 nucleosome upstream and the +1 nucleosome downstream of the TSS. Finally, downstream of the +1 nucleosome a regularly spaced nucleosome array is present, whose nucleosomes are called +2, +3 and so on (Figure 1.B).

# Figure 1



**Figure 1: MNase digestion flowchart and digestion level bias.** (A) MNase digestion of the chromatin produces oligo-, mono- and sub-nucleosomes which proportion is dependent on the digestion level. On the right a scheme of the chromatin MNase digestion is depicted. Black arrows represent the fraction generation from chromatin or upper fractions. Green arrows represent MNase pseudo-exo-nuclease activity whose fractions are subjected to. (B) Composite plot of smoothed and normalized nucleosome dyad frequency calculated from fragment occurrence around all BP promoters. Tracks are visualized in ascending order of digestion level. Despite nucleosome position that is preserved among digestion levels, nucleosome occupancy is strongly affected by MNase digestion level bias. Particularly, the -1 position upstream of the TSS shows strong sensitivity to MNase digestion.

### 1.3 Determinants of nucleosome mapping

The determination of nucleosome positioning and occupancy along the genome is a complex phenomenon that involves multiple factors. These can be divided in *cis*-factors, like the DNA sequence with its own biophysical properties, and *trans*-factors, such as chromatin remodelers, TFs, transcription machinery, histone PTMs, histone variants and other protein complexes.

On one side, a variation of the DNA sequence determines variations in its own biophysical properties, including the inclination to wrap around the histone octamer. On the other side, *trans*-factors play a relevant role in the determination of the nucleosome organization *in vivo*. In fact, they introduce the dynamism necessary for the organisms to regulate the chromatin structure and modulate gene expression to stimuli. The relative quantitative contribution of each factor to the nucleosome organization was amply debated in the last decade (Kaplan et al., 2010b; Zhang et al., 2010).

### 1.4 DNA sequence and nucleosome mapping

DNA is highly contorted by its assembly around nucleosomes. The bending around a histone octamer produces a 1.67 left-handed superhelical turn with around 80 bp per turn, mostly created by base pair roll in the minor and major grooves toward the histone octamer (Richmond and Davey, 2003). This tight wrapping has important consequences because a 147 bp DNA sequence with high flexibility is characterized by higher affinity to the histone octamer, leading to stronger interactions. In contrast, more stiff DNA sequences have a reduced affinity to the histone octamer, leading to higher accessibility to *trans*-factors and requiring a reduced amount of energy for remodeling. Indeed, it has been well known for decades that some sequences are thermodynamically preferred for nucleosome assembly than others (Drew and Travers, 1985). SELEX experiments demonstrated that some naturally occurring nucleosomal DNA sequences have several fold higher affinity to the histone octamer than the bulk genomic DNA (Thastrom et al., 1999). The same study also demonstrated that non-natural sequences, designed to follow biophysical rules for a higher bendability, have even 40 fold higher affinity to the histone octamer than the most favorable natural ones. This observation pointed to a DNA landscape

evolution that does not preclude accessibility into the nucleosomal DNA in absolute terms. Similar concepts were also found in genome-wide studies. In budding yeast, nucleosome maps obtained *in vitro* and *in vivo* are well correlated. Furthermore, these maps also correlated with a probabilistic model based on the preference of 5 bp long sequence oligomers to be packed in a nucleosome (Kaplan et al., 2009). However, they do not perfectly match the *in vivo* nucleosome positioning and organization. This indicates that DNA sequence alone cannot completely explain the nucleosome organization, pointing to a collaborative role played by *trans*-factors.

### 1.4.1 Di-nucleotide 10 bp periodicity

Most of the interactions between the nucleosomal DNA and the histone octamer are guided by residues located inside the minor groove (Luger et al., 1997). For that reason, the closer the minor grooves are to the histone octamer the stronger the interactions are. Nucleosomal DNA flexibility is boosted by A and T (WW) di-nucleotides 10 bp periodicity, located where narrowed minor grooves face inward toward the histone octamer (Satchwell et al., 1986). The highest contribution to the nucleosomal DNA flexibility is given by the TA di-nucleotide in these positions (Thastrom et al., 1999). Furthermore, the WW 10 bp periodicity can also directly affect the electrostatic interactions between arginine residues that penetrate in the narrowed minor groove (West et al., 2010). In parallel, G and C (SS) di-nucleotides 10 bp periodicity, out of phase to the WW periodicity, also increases the DNA bendability, in particular when the SS locates in the minor grooves that face outward from the histone octamer (Drew and Travers, 1985). This di-nucleotide organization can boost a higher DNA flexibility that can better accommodate the tension in the minor groove accumulated in the DNA wrapping around the histone octamer. Indeed, the WW 10 bp periodicity with a SS anti-periodicity was observed in the nucleosomal DNA of numerous species and using different nucleosome mapping genome-wide methods (Brogaard et al., 2012; Mavrich et al., 2008b). The 10 bp periodicity rule was also applied for designing strong nucleosome positioning sequences used in a plethora of *in vitro* studies (Lowary and Widom, 1998). The most commonly used nowadays is the Widom601. Moreover, the di-nucleotide 10bp periodicity can guide rotational positioning. For example, the positioning of some +1 nucleosomes and their further sliding downstream can preferentially follow WW 10 bp steps in the linker DNA between the +1 and the +2 nucleosomes (Cui et al., 2012).



### 1.4.2 Homopolymeric DNA sequences

Homopolymeric DNA traits affect nucleosome assembly. In *S.cerevisiae* and other species, promoters and transcription termination sites (TTSs) are strongly enriched for poly(dA:dT) traits compared to genomic average (Yuan et al., 2005), driving the formation of NDRs. A systematic manipulation of poly(dA:dT) traits in budding yeast promoters revealed a reduced nucleosome occupancy (Raveh-Sadka et al., 2012). How these traits drive nucleosome depletion is not completely known, but the most probable hypothesis points to its own biophysical properties characterized by high stiffness and low affinity to the histone octamer (Segal and Widom, 2009). Indeed, poly(dA:dT) sequences are less prone to be reconstituted into nucleosomes *in vitro* (Zhang et al., 2011) and to be packed in nucleosome *in vivo* (Suter et al., 2000). Nonetheless, a crystal structure of a nucleosome containing a 16 bp poly(dA:dT) trait was generated (Bao et al., 2006). Importantly, nucleosome depletion induced by poly(dA:dT) traits is not absolute, but rather quantitative (Segal and Widom, 2009). Most of *S.cerevisiae* genes are housekeeping, characterized by a constant transcription rate and low expression plasticity. Consequently, it is thought that poly(dA:dT) traits are evolutionary selected at their promoters in order to disfavor nucleosome assembly and facilitate the recruitment of the transcription machinery (Field et al., 2008). In parallel, *S.cerevisiae* promoters of stress-responsive genes do not contain these elements and they are characterized by higher intrinsic nucleosome occupancy and stronger dependency by trans-factors for their activation (Tirosh and Barkai, 2008). In several yeast species, also poly(dG:dC) traits were shown to have similar features (Tsankov et al., 2011). In conclusion, homopolymeric traits play a predominant role in several yeast species for the NDR formation and consequently in the nucleosome organization around promoters. On the other hand, this sequence signature is not universal. Other yeast species, such as *S.pombe* (Lantermann et al., 2010), promoters are not significantly enriched for poly(dA:dT). Interestingly, no extensive study in *D.melanogaster* has been performed in this regard.

### 1.4.3 G + C content

It is well known in literature that G+C rich sequences well correlate with nucleosome occupancy *in vitro* (Tillo and Hughes, 2009). In this dissertation, I define the sum of the relative percentage for G and C nucleotides as GC content. Low GC content in *S.cerevisiae* and *D.melanogaster* promoters was correlated to nucleosome depletion and high GC content at mammalian promoters

was associated to high intrinsic nucleosome occupancy *in vitro* (Valouev et al., 2011). Moreover, a model based on the GC di-nucleotide can partially predicts the +1 nucleosome location in *D.melanogaster* (Mavrich et al., 2008b). The correlation between GC content and nucleosome occupancy can be explained through many structural properties of the DNA, and/or by reduced frequencies of poly(dA:dT) traits (Tillo and Hughes, 2009). Firstly, measurements of nucleosome stability and nucleosomal DNA distortion *in vitro* showed that GC-rich sequences better accommodate DNA distortions induced by DNA wrapping around the histone octamer. The same study also showed that nucleosome stability increases with higher GC content with the tested limit of 55-57% (Chua et al., 2012). In contrast, linker DNAs are enriched in AT content. In addition, 5mers composed exclusively of A and T nucleotides correlated with the lowest nucleosome occupancy (Kaplan et al., 2009). Nevertheless, the relationship between GC content and nucleosome occupancy is not absolute and linear. As abovementioned, long poly(dG:dC) traits disfavor nucleosome packaging. In addition, the most common 5mers around the nucleosome dyad, which is the region with the strongest interactions between histone and DNA, are characterized by an average of 1.4-1.8 A or T bases in humans. This AT frequency corresponds to a slight reduction against the genomic average of 2.5 (Prendergast and Semple, 2011). Furthermore, local variations of GC content can drive functional differences in nucleosome occupancy. For example in many species exons and introns have a different GC content (exons: 48%; introns: 40% in *D.melanogaster*) which correlates with a concomitant variation in nucleosome occupancy.

## 1.5 Nucleosome motions and spontaneous unwrapping

Nucleosomes are not static entities, but rather subjected to spontaneous thermodynamic fluctuations in their structure that can affect accessibility to the nucleosomal DNA. Those include nucleosome breathing and its spontaneous unwrapping, spontaneous nucleosome sliding (movement of a nucleosome in other close positions), dissociation and partial assembly of the histone octamer, and nucleosome gapping (Eslami-Mossallam et al., 2016). In this introduction, due to their relevance in this work, only nucleosome breathing and spontaneous unwrapping are explained in more detail.

Nucleosome breathing is the mechanism in which a section of DNA at the entry/exit of the nucleosome unwraps (called spontaneous unwrapping), partially relieving the mechanical stress. DNA sequence and histone tail PTMs were shown to play a major and minor role in this mechanism, respectively (Anderson and Widom, 2000; Polach et al., 2000). Evidences about this phenomenon were produced by restriction enzymatic accessibility assays (Polach and Widom, 1995) and single-pair Forster resonance energy transfer (FRET) experiments (Koopmans et al., 2008). The former showed that restriction enzymes activity was high at the entry/exit of nucleosome and was strongly reduced toward the nucleosome dyad. Identical observations, but with a more direct approach, were obtained with FRET experiments, by attaching donor and acceptor dyes at different coordinates within the nucleosome. In a wrapped conformation, FRET signal was higher since donor and acceptor dyes were close in space. In contrast, in an unwrapped conformation, the signal decreased due to the distance between the dyes. Both approaches suggested a step-wise unwrapping mechanism that starts from the entry/exit of the nucleosome and proceeds toward the dyad (Polach and Widom, 1996).

These results agreed with a near bp resolution map of the strength of the DNA-histone interactions produced by mechanical unzipping of single nucleosomes (Hall et al., 2009). The interaction map was characterized by three regions of strong interactions located at the dyad and at +/- 40 bp from it. Conversely, the interaction strength at the entry/exit of the nucleosome was confirmed as weak. Interestingly, the analyzed nucleosomes showed a polarity in resistance to the unzipping, since a different interaction map with decreased overall interaction strength was generated by mechanical unzipping from the opposite side of the same nucleosome. Taking into account the symmetry of the histone folding between two nucleosome halves, the only source of difference that could explain the different interaction maps resided in the asymmetry of the underlying DNA sequence. This study used not only the Widom601 sequence, but a broader spectrum of DNA sequences, indicating the universality of their conclusions.

Spontaneous unwrapping is characterized by two important features that could drive the accessibility within the nucleosome: rapidity (Li et al., 2005) and asymmetry (Ngo et al., 2015). Firstly, Li and colleagues demonstrated that nucleosomal DNA remains wrapped around the histone octamer for only 250 ms before being subjected to spontaneous unwrapping. Secondly, Ngo and colleagues, through single molecule fluorescence force spectroscopy experiments, demonstrated that: 1) it mostly proceeds in an asymmetric fashion, involving one or the other

nucleosome side; 2) it leads to tighter DNA-histone interactions in the nucleosome half that remains wrapped. Importantly, asymmetric unwrapping was connected to the asymmetry of the DNA sequence between the two nucleosome halves, which in turn determines differences in DNA flexibility. In fact, asymmetric unwrapping mostly involves the half with less prominent TA 10 bp periodicity. To confirm this observation, they applied their method to a modified version of the Widom601 sequence containing identical TA 10 bp periodic steps between the two nucleosome halves. In this case the nucleosome side that underwent asymmetric unwrapping was stochastically determined.

## 1.6 Trans-factors and nucleosome mapping

Although DNA sequences can generate NDRs, influence nucleosome occupancy, affect rotational positioning and nucleosome motions, they cannot fully explain the nucleosome organization observed *in vivo*. The role of trans-factors in the determination of the nucleosome organization can be summarized in three main observations. Firstly, recent studies revealed inconsistencies in the statistical nucleosome positioning hypothesis. Secondly, some aspects of the nucleosome organization cannot be reconstituted with only purified histones and DNA (Salt Gradient Dialysis, SGD). Lastly, nucleosome models purely trained on the DNA sequence fail to predict nucleosome positions with high accuracy, including the -1 and +1 positions around the TSSs (Liu et al., 2014). Moreover, model accuracy around promoters tended to decline when applied to higher organisms, indicating the evolutionary relevance of *trans*-factors (Liu et al., 2014). Are examples of *trans*-factors: ATP-dependent chromatin remodelers, the transcription machinery and TFs.

Nucleosome maps are characterized by well-positioned nucleosomes around CREs and more fuzzy nucleosomes further from these elements. This observation is in accordance with the statistical nucleosome positioning hypothesis (Kornberg and Stryer, 1988). This model predicts well-positioned nucleosome arrays originated from a “boundary element” in which nucleosome spacing is solely governed by nucleosome density and steric hindrance. Boundary elements include proteins or complexes bound to the DNA and/or cis-factors like poly(dA:dT) tracts. Nonetheless, other experiments challenged this hypothesis, even if the concept of boundary

element still remains valid. Chromatin reconstitution experiments with low histone concentration (Zhang et al., 2011) and *in vivo* experiments conducted to reduce nucleosome density (Gossett and Lieb, 2012) showed that global nucleosome spacing was unchanged around CREs. This suggested active mechanisms that clamp nucleosomes together through the action of chromatin remodelers.

Nucleosome organization around CREs is a dynamic process and chromatin remodelers move nucleosomes around through ATP hydrolysis. Chromatin remodelers can cause nucleosome sliding, partial or complete nucleosome eviction, and nucleosome replacement or histone exchange with histone variants. Depending on the ATPase and subunit composition, each chromatin remodeler has a specific function. In *D.melanogaster* the chromatin remodeler complexes NURD, (P)BAP, INO80 and ISWI have unique genomic targets and produces distinct effects on the chromatin. Indeed, the first three increase nucleosome density, thereby stabilizing nucleosomes on disfavored sequences. Instead, ISWI complexes reduce nucleosome occupancy on favored sequences mostly by nucleosome sliding (Moshkin et al., 2012). Pugh and colleagues recreated the nucleosome organization around promoters *in vitro* through the addition of whole cellular extract and ATP to the SGD material. This result suggested the relevant role of chromatin remodelers in the determination of nucleosome positioning, occupancy and spacing *in vivo* (Zhang et al., 2011). In a follow-up study, the role of trans-factors in the nucleosome organization around CREs was further confirmed by Korber and colleagues by applying a genome-wide nucleosome reconstitution using purified single components. They showed the interplay among DNA sequences, histones, nucleosome organizing factors and mainly chromatin remodelers in the primary structure of the chromatin (Krietenstein et al., 2016).

The role of the RNA polymerases on the nucleosome organization around promoters is still debated. Transcription initiation seems to affect the fine-tuning positioning of the +1 nucleosome and transcription elongation seems to have a role in the positioning of the downstream nucleosomes, reviewed in (Struhl and Segal, 2013). In both cases is not clear if these observations directly derive from the transcription machinery or from the recruitment of chromatin remodelers.

It is well known that TFs can compete with histones for DNA binding (Workman and Kingston, 1992). Transcription factor binding sites (TFBSs) located inside a nucleosome are less accessible, and consequently TFs have to overcome this barrier. Differences in the ability of the

TFs to bind nucleosomal DNA can derive from protein abundance, ability to bind wrapped nucleosomal DNA, affinity in the recruitment of chromatin remodelers, and location of the TFBSs within the nucleosome. TFs able to bind nucleosomal wrapped DNA are called pioneer TFs (Zaret and Mango, 2016). Their binding on the nucleosomal DNA can open or close the chromatin through cooperativity with other TFs or by recruiting chromatin remodelers.

Finally, nucleosome mapping can be also dynamically regulated by other factors, such as introduction or removal of histone PTMs, histone assembly in non-canonical nucleosome structures (Khuong et al., 2015), and incorporation of histone variants.

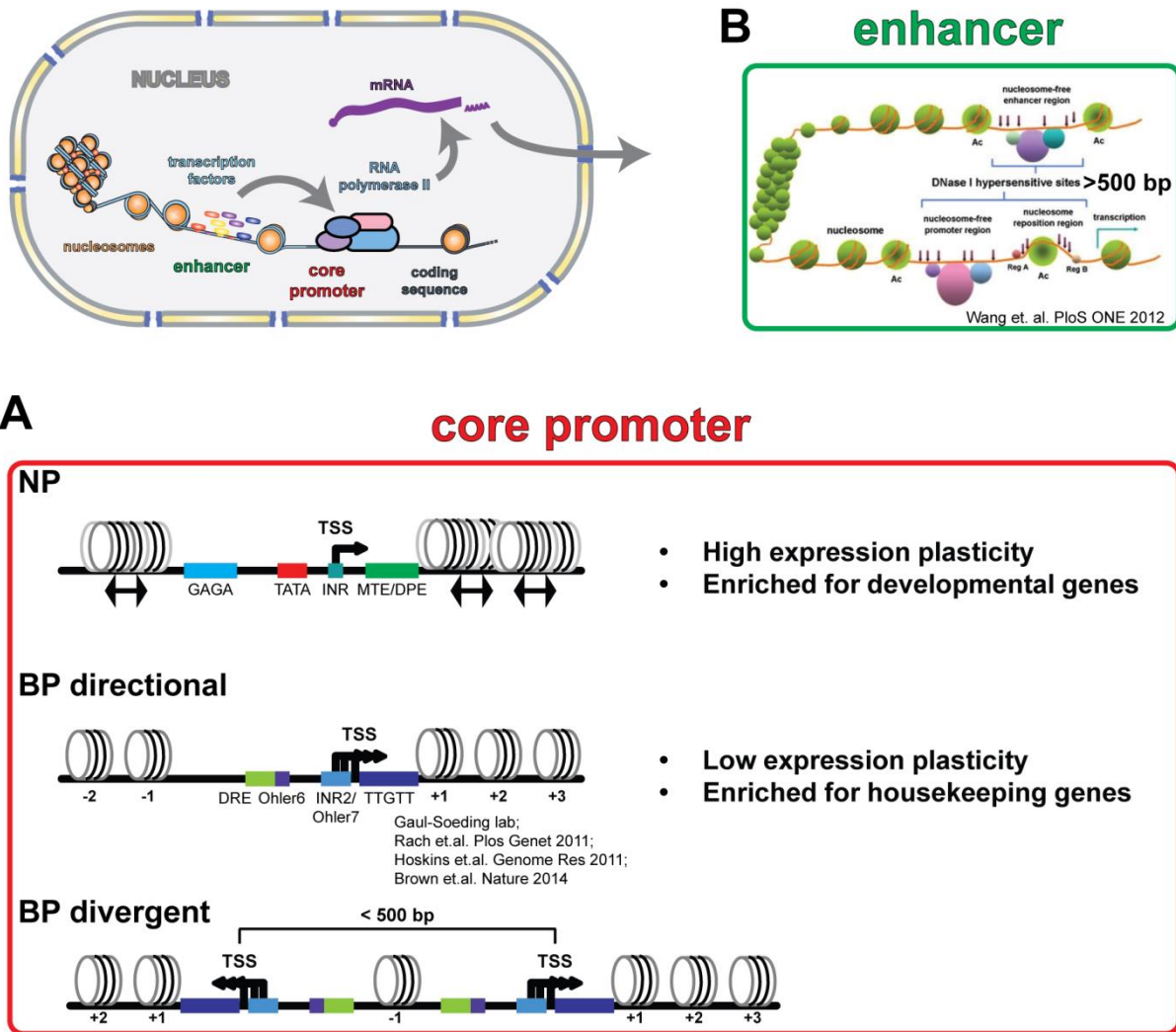
## **1.7 The core promoter and its shape**

The core promoter is defined as the minimal DNA sequence necessary for the recruitment of the transcription machinery to initiate transcription. A universal core promoter does not exist and each promoter shows its own peculiar DNA biophysical properties, motif compositions and chromatin features. This variability reflects distinct mechanisms of gene activation, RNA polymerases recruitment, and different strategies of engaged RNA polymerase pausing. Nevertheless, it is possible to distinguish core promoter architectures by clustering promoters with similar features and behaviors in response to upstream signaling. RNA polymerase II (Pol II) transcribes protein coding genes to produce messenger RNAs (mRNAs). Pol II recruitment requires general transcription factors (GTFs), such as TFIIB, TFIID, TFIIE, TFIIIF and TFIIH, forming the pre-initiation complex (PIC) (Sainsbury et al., 2015). For example, subunits of TFIIB and TFIID recognize distinct DNA motifs, including the TATA binding protein (TBP) and 13-14 TBP associated factors (TAFs) (Haberle and Lenhard, 2016). In parallel, the PIC subunit composition can change according to the promoter architecture (Sikorski and Buratowski, 2009).

To distinguish core promoter architectures, the spatial distribution of TSSs within the promoter is commonly used, namely the promoter shape. In this regard, it can be distinguished two promoter classes: narrow-peak (NPs) and broad-peak (BPs) promoters. The threshold used for the promoter shape distinction is arbitrary since the TSS distribution among promoters was revealed as a continuum of shapes. In this thesis, due to a general consensus in literature, NP

promoters are defined by a narrow distribution of TSSs, whereas BP promoters are defined by a broader distribution. In *D.melanogaster*, the promoter shape depends more on the underlying DNA sequence than trans-factors (Hoskins et al., 2011). NP promoters were mostly found in developmental and tissue specific genes, characterized by high expression plasticity (Graveley et al., 2011). They are also enriched for strongly positioned core promoter motifs, such as TATA box, Inr, MTE and DPE (Rach et al., 2009). On the other hand, BP promoters are mostly found in housekeeping genes and are enriched by broader distributed core promoter motifs, such as CpG islands (CGI) in vertebrates and Ohler 1, DRE, Ohler 6 and Ohler 7 in *D.melanogaster* (Hoskins et al., 2011; Rach et al., 2009). This dual promoter shape organization can biologically make sense if gene function is taken into consideration (Juven-Gershon and Kadonaga, 2010). Indeed, a sharp TSS distribution facilitates the control of Pol II recruitment and gene expression, which is necessary for developmental genes. In contrast, a broader TSS distribution reduces such control facilitating low expression plasticity typical of housekeeping genes. As confirmation, this distinction was also confirmed as an evolutionary standpoint, since two different degrees of conservation were found between the two promoter shapes. NP promoters show more evolutionary constrictions than BP promoters, since TSS mutations in the latter can be buffered by other initiation sites, leading to a small impact in the overall promoter strength (Schor et al., 2017). In addition, NP and BP shapes also show duality in nucleosome organization and Pol II pausing in human and *D.melanogaster*. Specifically, NP promoters were characterized by more fuzzy nucleosomes, less defined nucleosome organization and a sharper and more efficient Pol II pausing. Conversely, BP promoters were characterized by a typical canonical nucleosome pattern and a wider and less efficient Pol II pausing (Rach et al., 2011). The main features associated to NP and BP promoter classes are reported in Figure 2.A.

## Figure 2



**Figure 2: Cis-regulatory elements (CREs).** CREs are crucial in the regulation of gene expression. Transcription factors binding on distal regulatory elements, such as enhancers, mostly regulate when and where transcription takes place. In parallel, core promoters are central for the setting up of the expression level. (A) In this study, core promoters were also clustered by their promoter shape. Narrow-peak promoters (NP) are defined by a narrow distribution of transcriptional start sites (TSSs). These promoters are mostly found in developmental genes and enriched for strongly positioned motifs, such as TATA-box and Inr. The nucleosome organization around them is mostly characterized by fuzzy nucleosomes. Broad-peak promoters (BP) are defined by a broader distribution of TSSs. These promoters are mostly found in housekeeping genes and mostly show a canonical nucleosome pattern. In this study, BP genes were further divided in BP directional and BP divergent clusters. The latter included protein coding genes transcribed in opposite direction with a distance between TSSs of 200-500 bp. Therefore, the intergenic region between the two TSSs contains a shared -1 nucleosome. (B) In this study, enhancers were defined as distal DNase-I hypersensitive sites at least 500 bp distant from any TSS and transcriptional termination sites (TTSs).



## 1.8 Pol II pausing and nucleosome barrier

Between transcription initiation and the successive steps of productive elongation there is an intermediate one: promoter escape and promoter-proximal pausing. Pol II pausing is a highly regulated process during which the enzyme stalls in a region around 20-60 bp downstream of the TSS. The exact location of Pol II pausing is promoter- and specie-dependent. Numerous studies conducted in several model organisms demonstrated promoter-pausing as a widespread phenomenon and as an evolutionary conservative mechanism for gene regulation (Adelman and Lis, 2012). It was proposed that Pol II pausing serves as mechanism to synchronize expression of gene clusters involved in regulatory and developmental pathways (Liu et al., 2015).

Pol II mapping at base-pair resolution in *D.melanogaster* indicated that promoters with a focused initiation are characterized by a pausing with higher proximity to the TSS and higher efficiency, meaning high and sharp occupancy of the paused Pol II. Conversely, promoters with a broad initiation showed a wider and lower paused Pol II occupancy (Kwak et al., 2013). Promoters with highly efficient pausing are enriched for core promoter motifs, as well as for GAGA motif and its own binding protein: GAGA binding protein (GAF). For the less efficient pausing of BP promoters, a more prominent role played by the +1 nucleosome as barrier was postulated (Kwak et al., 2013). In regard of BP promoters, these findings were also confirmed and amplified by the Gilmour group. They found a correlation between less efficient pausing and a core promoter motifs (Motif 1) and its own binding protein: Motif1 binding protein (M1BP) (Fuda and Lis, 2013; Li and Gilmour, 2013).

In all cases, Pol II has to overcome the +1 nucleosome barrier in order to resume transcription. Indeed, the +1 nucleosome represents the main obstacle for transcription elongation, at least 2-3 folds higher than the +2 or further downstream nucleosomes (Weber et al., 2014). Several mechanisms were demonstrated to modulate the nucleosome barrier, including histone PTMs, histone variants replacement (H2A.Z and H3.3), histone chaperones and chromatin remodelers (Skene et al., 2014; Teves et al., 2014a). Interestingly, among these factors also the structure of DNA itself can play a role due to the bi-directional torsion forces, positive ahead and negative behind, generated by the action of Pol II. In fact, nucleosome destabilization mediated by the positive torsion was confirmed *in vivo* though accumulation of torsional stress induced by topoisomerases inhibition (Teves and Henikoff, 2014).

## 1.9 Probing genome accessibility

Open or accessible regions of the chromatin are hypersensitive to treatment with small and unspecific nucleases due to reduced nucleosome occupancy. These nucleases preferentially cut within protein-free DNA, like linker DNA, as well as in proximity or within DNA occupied by non-histone protein. Indeed, binding width and resident time of TFs, and the binding strength of transcription machinery on DNA, are inferior compared to the histone octamer, which thereby offers greater protection against nucleases. Among these nucleases, the most commonly used for probing chromatin structure are deoxyribonuclease I (DNase I) and MNase (Wu et al., 1979). DNase I and MNase enzymatic activities are used to directly measure accessible and protein protected regions of the chromatin, respectively. In the last decade, they have been coupled to next generation sequencing (NGS) technologies in order to sequence, identify and map the DNA fragments that survive their digestion in a genome-wide fashion (DNase-seq and MNase-seq).

### 1.9.1 DNase-seq and enhancers

DNase I is an unspecific double-strand endo-nuclease that preferentially cuts protein-free DNA. Nowadays, protocols of DNase-seq are conducted with an enzymatic limited digestion and size selection for short fragments to enrich for fragments that derive from double cutting events within NDRs (Vierstra et al., 2014). These regions are called DNase I hypersensitive sites (DHSs). This approach was used to reveal chromatin accessible landscapes of several species and cell types in order to identify active enhancer and promoter (Thurman et al., 2012). Alternative approaches were recently introduced for technical advantage: ATAC-seq (Buenrostro et al., 2013) and for a precise enhancer identification: STARR-seq (Arnold et al., 2013).

Nucleosome landscape around non-promoter DHS peaks (from now on called distal DHSs) is characterized by two well positioned nucleosomes that surround a NDR located within the peak. Depending on the TF dwell-time on the DNA, nucleosome arrays can also be present in both directions. This nucleosome organization is driven by three main factors: 1) variation of the underlying DNA sequence; 2) TF boundary effect; 3) recruitment of chromatin remodelers. DNA sequence plays a role both within the DHS peaks and in the surrounding regions. Indeed, TFBSs are generally located in regions with predicted high nucleosome occupancy due to the high GC content (Tillo et al., 2010). For that reason, TFs have to introduce relevant changes in the local

chromatin structure to access to their own binding sites. In binding sites occupied by (their cognate) TFs, flanking nucleosomes showed a more conserved occupancy and positioning (Gaffney et al., 2012). Therefore, activation of distal DHSs locally introduces strong variation in occupancy on nucleosomes that directly cover TFBSs, without drastic changes in the local nucleosome occupancy (West et al., 2014). Furthermore, the NDR within the DHS peak derives mostly by the competition between TFs and nucleosomes, which is affected by the affinity and residence time of the TFs (Vierstra et al., 2014). Indeed, surrounding nucleosome arrays are partially driven by a boundary effect introduced by the TF and are dependent on the TF residence time. Factors like the insulator CTCF are characterized by long residence time, indeed its binding on the DNA can lead to long well positioned nucleosome arrays composed of up to 20 nucleosomes (Fu et al., 2008). However, this configuration is an exception and nucleosome positioning and array formation around TFBSs were shown to be heterogeneous and asymmetric with regard to other TFs (Kundaje et al., 2012). Finally, chromatin remodelers also play a relevant role in the nucleosome organization around distal DHS peaks. Chromatin remodelers can introduce more profound NDRs at DHS peaks and can increase positioning of surrounding nucleosomes, as shown for the CTCF binding sites after chromatin remodeler knock-downs (Wiechens et al., 2016). Noteworthy, also the dynamic interactions between pioneer factors and chromatin remodelers are fundamental determinants for NDR formation during cell differentiation and development (Swinstead et al., 2016).

### **1.9.2 MNase-seq and nucleosome mapping**

Several techniques have been developed for nucleosome mapping: MNase-seq, chemical map (or CC-seq), MPE-seq, NucleoATAC-seq, NOMe-seq, and others. The first and most commonly used is MNase-seq. Micrococcal nuclease is an endo- pseudo-exo and calcium dependent nuclease that preferentially cuts protein-free DNA, such as linker DNA. Thus, histones and other DNA binding proteins protect the underlying DNA from the digestion. MNase preferentially digests single strand DNA. Indeed, it is thought that MNase cleaves one strand per time in close proximity, producing a double strand break. Successively and in a similar fashion, it chews the DNA until the enzyme reaches a protein obstacle, such as a nucleosome. The advantages of using MNase in nucleosome mapping compared to other nucleases are its higher resolution, due to the trimming of fragments ends close to the nucleosome borders, and a strong reduction in

activity when the enzyme meets a protein obstacle. MNase digestion of the chromatin releases several fractions: oligo-nucleosomes (particles constituted by two or more nucleosomes), mono-nucleosomes and sub-nucleosomes (originated from the digestion inside the nucleosome). The proportion among these fractions is strongly dependent on the digestion level. Commonly, a MNase-seq protocol consists of four main steps: 1) Nuclei isolation or cell permeabilization to facilitate the entrance of the enzyme into the nuclei. 2) A limited digestion in a calcium-containing buffer. This step is performed after protocol optimization through variation of the digestion time or titration of the enzyme concentration. A typical digestion level corresponds to a chromatin fragmentation with around 70-80% of the particles as mono-nucleosome (Rizzo et al., 2012). 3) Mono-nucleosomal DNA size-selection and isolation by cutting out the mono-nucleosome band from the agarose gel or by using magnetic beads. MNase digestion of the chromatin is complex and impossible to mimic computationally due to the numerous involved variables, as shown in Figure 1.A. Indeed, each fraction derives from upper ones and is subjected to MNase trimming in parallel. Moreover, nucleosome information can be lost due to the digesting away step. 4) Sequencing is preferentially performed by pair-end because it permits to recognize the location of both fragment ends. This leads to a more accurate calculation of the nucleosome dyad and a reduced percentage of fuzzy nucleosomes (Flores et al., 2014). Despite its advantages, MNase-seq has some limitations affecting both nucleosome occupancy and positioning. Nucleosome occupancy cannot be measured in absolute terms since it can be biased by the local chromatin accessibility, lost of NDR fragments, and strong dependency on the digestion level. Nucleosome positioning is affected by the MNase sequence bias, nucleosome fuzziness that occurs when a cell population is analyzed (Flores et al., 2014), and divergence between fragment ends and nucleosome borders. This divergence can derive from either digestion at the entry/exit of the nucleosomal DNA or from an incomplete digestion of linker DNAs (Nikitina et al., 2013).

MNase-seq is not only used for nucleosome mapping, but can also be applied for other purposes. As already mentioned, protection against MNase digestion is offered by any DNA-binding protein, a principle used in many studies conducted in the Henikoff lab by mapping small fragments (shorter than 75-50 bp) (Henikoff et al., 2011). Paused Pol II and TFs DNA protected fragments can be retrieved by MNase digestion as confirmed by chromatin immunoprecipitation (ChIP)-seq in *D.melanogaster* of Pol II (Teves and Henikoff, 2011) and for

some TFs in budding yeast and *D.melanogaster* (Kasinathan et al., 2014). On the other hand, MNase digestion can be used as chromatin shearing approach for ChIP-seq of chromatin-related features, such as histone PTMs, histone variants and less specific canonical histones. In this case, formaldehyde (FA) cross-linking can be optionally used before chromatin preparation, since FA bridges could better preserve DNA-histone interactions avoiding nucleosome re-arrangements during sample preparation and digestion (Zhang and Pugh, 2011). The advantages of a MNase-ChIP-seq are multiple: single-nucleosome resolution for epigenetic marks mapping; possibility to integrate information between nucleosome positioning and their PTMs (Weiner et al., 2015); confirmation that nucleosome positions obtained through MNase-seq are indeed generated from nucleosomal particles (Chereji et al., 2016).

Nucleosome mapping with MNase-seq was performed in a numerous amount of studies (Teif, 2016) and in different organisms, including budding yeast (Lee et al., 2007), *D.melanogaster* (Mavrich et al., 2008b), human cell lines (Valouev et al., 2011) and many other species. These studies showed a strong evolutionary conservation of the nucleosome organization around CREs, such as the presence of NDRs around TSSs, TTSs and distal DHSs, a canonical nucleosome pattern around the TSS of many genes, and well-positioned nucleosomes that surround TTSs and distal DHSs. Furthermore, these studies showed more pronounced nucleosome depletion around promoters and increased positioning of the surrounding nucleosomes in expressed genes compared to repressed ones, which are characterized by more occupied NDRs and increased nucleosome fuzziness (Valouev et al., 2011). However, differences among species were also found, such as the NDR width around TSSs, the +1 nucleosome position, and the frequency of distinct nucleosome patterns (Mavrich et al., 2008a; Mavrich et al., 2008b). These differences could derive from divergences in transcriptional regulatory mechanisms, *trans*-factor content and genome complexity.

To overcome some of MNase-seq limitations, an innovative method was developed for nucleosome mapping, called chemical map or CC-seq. This method has been applied in *S.cerevisiae*, *S.Pombe* and mammalian cells (Brogaard et al., 2012; Moyle-Heyrman et al., 2013; Voong et al., 2016) and can provide nucleosome dyad mapping at single base pair resolution. Chemical maps showed great consistency with the MNase ones, but also revealed a stronger 10 bp periodicity of the nucleosomal DNA, indicating the relevance of this sequence feature for the fine-tuning of nucleosome positioning. It is likely that the magnitude of the 10 bp periodicity was

underestimated in MNase-seq maps due to imprecise correspondence between sequenced fragment ends and nucleosome borders derived by a poor alignment of nucleosomal DNAs in meta-plots. On the other hand, chemical mapping also owns technical limitations and is not immune from biases. Indeed, it requires an engineered H4S47C to function, which is technically difficult to obtain in organisms with multiple copies of the H4 histone gene. Moreover, it also requires higher depth of sequencing compared to MNase-seq in order to boost the nucleosome signal from the noise, which could also derive from unspecific labelling of non-histone proteins containing cysteines.

## 1.10 MNase-seq biases

MNase-seq is affected by biases: 1) sequence preference of cutting; 2) ability to digest inside the nucleosome; 3) dependency on the digestion level.

It has been known for decades that MNase preferentially cuts DNA with an A or T at the 5'. Digestion of end-labelled linear DNA molecules with known sequence revealed that more than 95% of the cut sites contain an A or T at the 5'. In fact, it was measured that MNase rate of cleavage within these nucleotides is at least 30 folds higher than G or C nucleotides (Dingwall et al., 1981). Furthermore, the same study demonstrated the TA di-nucleotide as the preferred cut site, also more frequent than the complementary AT. It is important to make a distinction between cutting preference at TA di-nucleotides and within AT-rich sequences. Indeed, similar experiments showed that MNase does not preferentially cuts within sequences composed only of A and T, but it rather prefers to cut within a more heterologous sequence environment (Horz and Altenburger, 1981). Particularly, TA di-nucleotides surrounded by G and C are the most preferred site for MNase cutting (Dingwall et al., 1981; Horz and Altenburger, 1981).

The doubt that MNase-based nucleosome maps are strongly determined by sequence bias and size selection was also pointed out. In budding yeast, a high correlation was observed among *in vivo* and *in vitro* nucleosome maps, GC content profile and digestion on genomic naked DNA (gDNA), both at genome-wide and single gene levels (Chung et al., 2010). At the genome-wide these correlations should be taken carefully, because models trained on naked DNA digested with MNase or disrupted with mechanical sonication can also partially predict nucleosome

occupancy (Locke et al., 2010). Moreover, shearing through sonication also preferentially disrupts the chromatin in AT rich regions, although to a lesser extent than MNase cuts. At the same time, the sequence bias can be mitigated by using a gDNA control. Nonetheless, no significant change in the nucleosome meta-profiles around TSSs and TTSs was observed after gDNA correction, pointing to a mild effect played by the sequence bias in the data (Deniz et al., 2011). At single gene level, the reliability of MNase-seq data is more controversial at some positions. For instance, Chung and colleagues pointed to a nucleosome position in the intergenic region between Gal1 and Gal10 genes, suggesting that the signal was an artifact driven by the local higher GC content. However, another study showed the same position as a binding site for the RSC complex (a Swi/Snf chromatin remodeler) with a partially unwound nucleosome, a configuration that facilitates nucleosome eviction. Indeed, in absence of RSC the same position is occupied by a canonical nucleosome (Floer et al., 2010). At the same time, the NDR at the TTS was also proposed as an artifact of MNase sequence bias due to its enrichment for poly(dA:dT) traits (Fan et al., 2010), but a substantial depletion was later confirmed through chemical mapping (Brogaard et al., 2012). Finally, to confirm that nucleosome mapping was not exclusively driven by its sequence bias, chromatin was prepared *in vitro* with histones and DNA from different species and cut with MNase and caspase-activated DNase. The latter is characterized by a different sequence bias and is not able to cut within nucleosomes due to steric hindrance. The two approaches showed pretty high correlation, pointing to the simpler notion that MNase-seq experiments mapped indeed nucleosomes (Allan et al., 2012).

In conclusion, MNase has a strong preference of cutting for an A or T at the 5'. However, most of the studies that characterize MNase cut sites were restricted by a limited number of analyzed sequences. Thus, a better characterization of the sequence content around MNase cutting sites in a genome-wide fashion is required. MNase sequence bias can be studied and reduced at genome-wide by performing a gDNA control, as demonstrated in yeast. Nevertheless, these studies based their conclusions mostly on genome-wide correlations. Hence, a detailed characterization of the sequence bias around CREs in *D.melanogaster* is also required.

In regard of the intra-nucleosome digestion, little information is available. MNase digestion of the chromatin also releases sub-nucleosome particles due to the small size of the enzyme compared to the nucleosome (Allan et al., 2012). Moreover, transient accessibility following spontaneous unwrapping or active ATP-consuming mechanisms can also facilitate the

nucleosomal DNA digestion. MNase primary cuts derived from extensive digestion conditions showed that single strand cuts are more frequently followed by a second cut on the other strand, producing a double strand break, rather on the same strand (Cockell et al., 1983). This study was conducted on a limited set of sequences. Hence, a more profound characterization on how MNase cuts within the nucleosome at genome-wide level is necessary.

Furthermore, an additional MNase-seq bias is the dependency on the digestion level. Indeed, different nucleosome datasets can be generated from the same biological sample according to the applied experimental conditions, as shown from different nucleosome meta-profiles around CREs generated from the same material, but from different laboratories (Ho et al., 2014). This discrepancy mostly derives from: 1) different saline concentration used for chromatin extraction; 2) different digestion level. Firstly, low salt extracted chromatin (80-150mM NaCl) was shown to be enriched for less stable and more soluble nucleosomes, mainly containing histone variants such as H2A.Z (H2A.v in *D.melanogaster*) and H3.3 (Henikoff et al., 2009). These nucleosomes are mostly located in active chromatin regions. Secondly, Friedman and colleagues reported that nucleosomes can be differently susceptible to MNase digestion according to the digestion level. They introduced the concept of “differential MNase-seq” by sequencing mono-nucleosome bands of several digestion levels obtained through MNase titration (Weiner et al., 2010). Specifically, they analyzed three digestion levels in budding yeast that were defined by the ratio in DNA signal between the mono-nucleosome band versus the rest of the digested chromatin: under-digestion (15% mono-nucleosomes); typical-digestion (80% mono-nucleosome) and over-digestion (100% mono-nucleosomes). By comparing the three digestion levels around TSSs and TTSs, Friedman and colleagues found a great overlap in nucleosome positioning, but profound differences in nucleosome occupancy mostly between under-digestion and the other two digestion levels. Particularly, nucleosome occupancy at the -1 position upstream of the TSS and at positions surrounding the TTSs was strongly affected by the digestion level bias. Consequently, a direct comparison among nucleosome maps of different studies can be performed only when characterized by similar digestion levels (Rizzo et al., 2012). The discovery of the digestion level bias provided new insights in the chromatin field, showing the susceptibility of some nucleosome to MNase digestion and affecting the validity of previous studies in which typical- and over-digestion levels were mostly used. Thus, a short-digestion level is now required as control as well as can be considered as a new tool to study



nucleosome organization. In this dissertation, a condition of under-digestion or incomplete digestion of the chromatin is defined as short-digestion. The most common digestion level applied in previous studies, which is an intermediate digestion level, is defined as typical-digestion. Finally, over-digestion or complete digestion of the chromatin into mono-nucleosomes is defined as long-digestion.

In conclusion, digestion level bias can strongly affect nucleosome occupancy, but how it is generated throughout a differential MNase-seq is not fully elucidated. Particularly, it is missing a characterization of what information oligo- and sub-nucleosomes contain among digestion levels. Indeed, these two fractions were not considered or not analyzed separately in previous studies that apply a differential MNase-seq.

### **1.11 Differential MNase-seq (Fragile/Resistant nucleosomes)**

More recently, the digestion level bias has been used as differential MNase-seq in order to characterize novel features of the nucleosome organization. Xi and colleagues performed the very first genome-wide analysis of the chromatin landscape applying a differential MNase-seq in budding yeast. Their purpose was the identification of nucleosome positions susceptible to MNase digestion (Xi et al., 2011). These nucleosomes were called fragile, speculating the existence of biophysical properties behind their sensitivity to MNase. In this regard, it is important to distinguish the terms MNase-sensitivity and fragile or resistant nucleosomes. MNase-sensitivity is just a metric that indicates how easily each nucleosome is digested by MNase. Conversely, nucleosome fragility and resistance have a more pronounced biological meaning by including in their definitions biophysical properties of the DNA-histone interactions that drive MNase-sensitivity, such as underlying DNA sequence and active mechanisms. MNase-sensitive and fragile nucleosomes are characterized by a reduction in occupancy with the increase of the digestion levels. In contrast, MNase-insensitive and resistant nucleosomes are characterized by higher occupancy with the increase of the digestion level (Chereji et al., 2016; Mieczkowski et al., 2016). To identify fragile nucleosomes, Xi and colleagues measured the ratio in occupancy between short- and long-digestion levels for each called nucleosome. Fragile nucleosomes were defined with occupancy at least three fold higher in the short versus the long

digestion. In this way, they revealed the relative abundance of fragile nucleosomes along the genome, which was around 5% among called positions. Most importantly, they showed their enrichment in active genomic regions including upstream of the TSSs, around the TTSs, within tRNA gene bodies, within replication origins and in some transposon elements. Lately, enrichment of MNase-sensitive nucleosomes within and around DHSs was also described in *D.melanogaster* (Chereji et al., 2016), although without a distinction between promoter and distal DHSs. Finally, Xi and colleagues also observed fragile nucleosomes in the promoter NDR, specifically in around 5% of Pol II transcribed genes. This position was afterward called 0, in order to distinguish it from the -1 and +1. The existence of the 0 nucleosome for certain genes is still matter of debate (Kubik et al., 2017a). Indeed, it was confirmed in ChIP-seq experiments of histones and histone variants, both in *D.melanogaster* and human (Chereji et al., 2016; Henikoff et al., 2009; Jin et al., 2009), but it was disproved in budding yeast by the same typology of experiments (Chereji et al., 2017; Rhee et al., 2014). Moreover, the nature of the 0 nucleosome is also not clear. On one hand, a non-canonical structure was suggested since enrichment for subnucleosomal histone-containing particles was observed by using a MNase independent method (Ishii et al., 2015). On the other hand, a recent study questioned its and the overall existence of MNase-sensitive nucleosomes at the TSS in *S.cerevisiae*, suggesting their origins as DNA fragments protected by non-histone barriers (Chereji et al., 2017) This controversy can be partially explained by a nomenclature issue since some MNase-sensitive nucleosomes that appear in the 0 position can be considered as -1 at single genes level, due to TSS miscalling or narrow NDRs. With regard to the fragile -1 nucleosome, Kubik and colleagues clustered budding yeast promoters in accordance to nucleosome organization and to the -1 nucleosome fragility produced by differential MNase-seq. They demonstrated a NDR width larger than 150 bp as prerequisite for fitting in a -1 fragile nucleosome (Kubik et al., 2015). In contrast, shorter NDRs remained mostly unoccupied due to steric hindrance. Interestingly, in these cases, the NDR genesis was mainly due to the chromatin remodeler RSC and general regulatory factors. Furthermore, Kubik and colleagues revealed that stable and fragile -1 nucleosomes have distinct sub-nucleosome fragment distributions even in short digestion levels. Fragile -1 nucleosomes showed a minimal protection of 100 bp, whereas stable -1 nucleosomes of 140 bp, suggesting that the digestion inside the nucleosomes can partially explain MNase sensitivity. Despite the debate regarding the 0 position, broader consensus on MNase-sensitivity along the chromatin

exists, at least in high eukaryotes. However, studies from budding yeast have to be taken carefully, since the detection of histones in MNase-sensitive nucleosomes is more ambiguous compared to other species for still unclear reasons. Finally, MNase-sensitive nucleosomes around CREs were also described in other species, including worm and plants (Jeffers and Lieb, 2017; Vera et al., 2014), and by using different experimental approaches, such as chemical mapping in *S.Pombe* and human cell lines (Moyle-Heyrman et al., 2013; Voong et al., 2016).

Genesis of fragile and resistant nucleosomes is still not clear and matter of debate, but it is most likely a combination of multiple factors. A non-histone origins and MNase sequence bias have to be considered as well, since most of MNase-sensitive nucleosomes are located in AT-rich regions. Nonetheless, Xi and colleagues investigated histone variant content within fragile nucleosome in budding yeast, revealing enrichment for the histone H2A.Z (Xi et al., 2011). This is in agreement with data on histone replacement with the H2A.Z/H3.3 dimer, which can generate nucleosomes that are less stable in solution in human (Jin and Felsenfeld, 2007) and are easily recovered with low-salt preparation of the chromatin (Henikoff et al., 2009). However, these observations cannot totally explain the nature of fragile nucleosomes. Firstly, H3.3 is not present in budding yeast and replacement only with H2A.Z in human does not increase nucleosome solubility. Secondly, cross-linking experiments led to a reduced but not complete absence of nucleosome fragility (Jin et al., 2009). Furthermore, the role of the DNA sequence was also analyzed, showing an increased GC content between mono-nucleosome fragments from a short- to a long-digestion level in *D.melanogaster* (Chereji et al., 2016), as well as local enrichment of poly-A and GC-rich motif in budding yeast (Kubik et al., 2015). Additionally, the role of trans-factors was also investigated. In a specific case, the recruitment of GRF and RSC complexes was shown to enlarge the NDR, creating the condition for a nucleosome to fit in and further destabilizing the fragile nucleosomes in a relevant number of cases (Kubik et al., 2015). Finally, a role for nucleosome motions and chromatin remodelers was also hypothesized (Kubik et al., 2017b). Mieczkowski and colleagues applied a MNase titration with four digestion levels and mild size-selection (fragments up to 1 Kb) to defined a score called MACC (MNase accessibility) (Mieczkowski et al., 2016). In each 300-500 bp genomic bin they computed the regression slope of linear fitting using as points the fragment frequencies among digestion levels. In this fashion, MACC measures how much each genomic bin is protected against MNase. They applied this method both in fly, mouse and human cell lines with and without canonical histone

immunoprecipitation. Their meta-profiles were consistent among datasets, including the presence of MNase-sensitive nucleosomes upstream of the TSS. Most importantly, they demonstrated that nucleosome occupancy and DNA accessibility do not always anti-correlate since some regulatory regions were characterized by simultaneous high nucleosome occupancy and DNA accessibility. These regions were enriched for many histone PTMs and variants typical of active chromatin regions. Same observations were also reported during chromatin changes around CREs in acute transcriptional inductions in *D.melanogaster* (Mueller et al., 2017). Additionally, also TFs can play a role in determining MNase sensitivity. Indeed, pioneer factors FoxA1 and FoxA2 were reported to maintain higher accessibility on tissue-specific enhancers through linker histone H1 disposal in mice, leading to the presence of MNase-sensitive nucleosomes (Iwafuchi-Doi et al., 2016).

In conclusion, differential MNase-seq is a useful tool to study nucleosome organization by assessing MNase sensitivity and/or nucleosome fragility and resistance. Previous studies have taken into consideration several factors to explain differential MNase sensitivity, but none of them considered sensitivity to MNase within individual nucleosomes, which presumably derives from the strength in DNA-histone binding. In addition, MNase sensitivity around CREs of *D.melanogaster* has been studied without considering promoter architecture and how it changes according to variations of the underlying DNA sequence.

## 2 AIM OF THE THESIS

Nucleosome organization around promoters and enhancers as well as the features that drive it are key questions in the field of gene regulation. MNase-seq is still the method of choice for nucleosome mapping, due to its simplicity and high resolution.

As described above, MNase digestion is affected by biases. Firstly, MNase has a strong preference for cutting at the A or T. However, a detailed characterization of the sequence content around MNase cut sites genome-wide is still missing, as well as a detailed investigation of the sequence bias in *D.melanogaster*, specifically around CREs. Secondly, MNase is able to cut within the nucleosomal DNA. In this regard, little information is available on how this occurs, and a characterization of the subnucleosomal digestion is also useful for a better understanding of MNase sensitivity.

To address these questions, **PART I** focuses on the MNase digestion to investigate its sequence bias at the cut sites and around CREs, and to characterize how MNase cuts within the nucleosome.

Nucleosome maps are affected by the digestion level and the mono-nucleosome size-selection. In this regard, it is not well understood what information is contained in the oligo- and sub-nucleosome fractions and how this information can be integrated within differential MNase-seq. Nucleosomes show a differential sensitivity to MNase digestion: some nucleosomes are more susceptible to MNase digestion and thus MNase-sensitive or fragile, while others are more insensitive or resistant to digestion. To study MNase sensitivity, previous studies focused mostly on the general chromatin accessibility, but little effort has been made to study how MNase sensitivity is affected by the DNA-histone binding strength within individual nucleosomes, which might more strongly influence nucleosome fragility and resistance. Moreover, in *D.melanogaster*, MNase sensitivity was studied around promoters only by considering gene expression level, therefore without taking into consideration gene functionality, which is partially captured by the promoter shape distinction. Finally, a deeper investigation on how nucleosome fragility and resistance are connected to sequence and activity components is

desirable, in order to achieve a better understanding of the nucleosome organization around CREs.

To address these questions, **PART II** develops a new differential MNase-seq approach to study the chromatin landscape around CREs, by analyzing separately the information contained in oligo-, mono- and sub-nucleosomes over different digestion times. Moreover, the relationship between nucleosome organization, fragile and resistant landscapes on the one hand and sequence features, such as the 10 bp di-nucleotide periodicity and GC content, and activity components is investigated. For this purpose, differential MNase-seq was integrated with promoter shape (CAGE), expression level (DTA-seq), chromatin accessibility (DNase-seq), Pol II occupancy (ChIP-seq) and gene organization data.

Finally, **PART III** measures how chromatin-related features can be affected by the MNase digestion level bias. A MNase short-digestion protocol is used as a chromatin shearing approach to preserve MNase-sensitive nucleosomes in the starting material. Four histone tail PTMs enriched in active chromatin regions are used to test this approach (ChIP-seq).

## 3 METHODS

### 3.1 Cell biology, molecular biology and biochemical procedures

#### Cell culture

*D.melanogaster* S2 cells were cultured in synthetic, serum-free Express Five medium (Gibco, 10486-025). 1 liter of Express Five medium was supplemented with 90 ml of 200 mM L-Glutamine (Gibco, 25030-081). Cells were thawed at passage 13 and cultivated until passage 20. During cultivation cells were grown at 25°C without CO<sub>2</sub> as semi-adherent monolayer in tissue culture flasks (Corning). When 90% confluent, cells were split into fresh flasks by means of seeding  $0.8 \times 10^6$  cells/ml. Cell counting and assessment of cell viability were performed using the Cell Counter and Analyzer System (CASY, Roche).

S2 cells were harvested by centrifugation (500 g, 4°C, 5 min), then washed with 2.5 ml ice-cold 1x PBS and centrifuged again with the same parameters. Cell pellets were kept and supernatant discarded. If pellets were not directly utilized for experiments, they were shock frozen in liquid nitrogen and then stored at -80°C. The amount of initial cells to generate cell pellets was calculated according to the subsequent experiment.

#### Cell formaldehyde cross-linking / de-cross-linking

Cross-linking was performed with  $50 \times 10^6$  S2 cells in a volume of 15 ml, with methanol-free formaldehyde (Polyscience, 18814) at the final concentration of 1% for 10 minutes. 125 mM Glycine was used as quencher, and then the cell pellet was washed in D-PBS. MNase digestion protocol was applied on  $25 \times 10^6$  cross-linked S2 cells following the same protocol applied for native chromatin. For all experiments de-cross-linking was performed through incubation of the sample at least for 3hr to a maximum of overnight (o.n.) at 65°C. This step reversed the cross-linking bridges of FA molecules.

#### Genomic DNA extraction

Genomic DNA (gDNA) was extracted from  $25 \times 10^6$  S2 cells pellets. Briefly, cells were resuspended in 600 µl of Nuclei Lysis buffer (10 mM Tris pH 8, 400 mM NaCl, 2 mM EDTA)

and then 200 µl of 1 mg/ml Proteinase K (Sigma-Aldrich) and 80 µl of 10% SDS were added. Subsequently, samples were incubated at 55°C overnight. To disrupt DNA-protein interactions, 460 µl of 6 M NaCl was added and then samples were centrifuged (11000 g, 30 min). The supernatant was recovered and centrifuged again (11000 g, 10 min). For DNA precipitation, 1 volume of 100% ethanol was added and samples were stored at -20°C for at least 1 hour. Successively, samples were centrifuged (11000 g, 4°C, 30 min) and the pellets were washed with 500 µl of 70% ethanol. Subsequently, pellets were air dried and the gDNA was resuspended in 75 µl of 0.1x TE buffer. Finally, 4 µl of RNase cocktail (Ambion, AM2286) was added followed by an incubation at 37°C for 30 min.

### **Alkaline agarose gel**

To generate the positive control, 2 µg of each oligos were annealed in 20µl of annealing buffer (10 mM Tris-HCl pH 7.4, 1 mM MgCl<sub>2</sub>, 100 mM NaCl). The solution was warmed up to 99°C for 10 min and cooled down to room temperature. The dsDNA product and oligos were analyzed by a 3% TAE agarose gel. 500 ng of annealed oligos were loaded in each gel. Alkaline agarose gel and running buffer were prepared with 0.1 volume of 10x buffer A (26.4 ml NaOH 50%, 20 ml EDTA 0.5 M, 953.5 ml H<sub>2</sub>O) without intercalant. DNA samples were purified by ethanol precipitation and dissolved in 20 µl of 1x buffer A and 0.2 volume of 6x buffer B (3ml NaOH 1M, 12 µl EDTA 0.5M, 1.8 g Ficoll 400, 0.015 g bromocresol green, 0.025 g xylene cyalon, H<sub>2</sub>O until 10 ml). DNA ladder was mixed 1:1 with a solution 4:1 of buffer A/buffer B. Electrophoresis was performed at < 3.5 V/cm. After running the gel was soaked in a neutralization solution (60.57 g Tris-HCl pH 7.6, 43.83 g NaCl, H<sub>2</sub>O until 500 ml) and it was stained with 0.5 µg/ml of GelRed (Biotium, 41001) in 1X TAE.

### **Immunoblot and peptide arrays**

Protein samples were mixed with 1x NuPAGE LDS Sample buffer (ThermoFisher, NP0007) and 0.1 M DTT and incubated at 80°C for 10 min. Lysate preparations directly from cell was used to evaluate antibody specificity for histone tail PTMs and rpb3. Pellets of 1 x 10<sup>6</sup> S2 cells were lysed with 300 µl of cOmplete Lysis-M buffer (Roche) implemented with cOmplete Protease Inhibitor cocktail (Roche). Lysed cells were then incubated on ice for 10 min and centrifuged (14.000 rpm, 4°C, 10 min). Protein-containing supernatants were recovered. Protein concentration was assessed with the Pierce BCA Protein Assay kit (ThermoFisher, 23225) according to manufacturer's instructions. Western blot was performed following standard



procedures. Protein separation was performed by SDS-PAGE electrophoresis. After run, proteins were transferred to a nitrocellulose membrane in transfer buffer for 1 hour at 300 mA. Membranes were blocked with 5% milk in TBST for 2 hours. Primary antibodies were diluted in blocking solution and incubated at 4°C overnight. On the next day, washes were performed with TBST. Subsequently, secondary antibodies were diluted in blocking solution and incubated for 1 hour. Subsequently, additional washes were performed with TBST, followed by a final wash in 1x PBS. Chemiluminescence was triggered with the Amersham ECL Prime Western Blotting Detection Reagent kit (GE Healthcare, RPN2232), and signal detected on films. To check histone tail PTMs antibody specificity and cross-reactivity commercialized peptide arrays were used, in which 89 unmodified and modified peptides from all canonical histones were immobilized (Millipore, AbSurance, 16-665 and 16-667). Array preparation and immunoblot were performed according to manufacturer's instructions. Rpb3 homemade antibody working concentration was 1:1000. For the histone tail PTMs, working solutions used for the peptide arrays were: H3K4me3 (0.1 µg/ml); H3K9ac (0.4 µg/ml); H3K18ac (0.1 µg/ml); H3K27ac (0.1µg/ml).

### **qPCR for ChIP evaluation**

qPCR was performed as quality control for the ChIP-seq protocol prior library preparation. A dilution of 1:10 of the IPs and 1% inputs were used to set up a 10 µl PCR reaction containing 5 µl of 2x SSOFast EvaGreen Supermix (Bio-Rad, 172520) and 0.3 µl of both forward and reverse 20 µM primers. PCR was performed in a Bio-Rad CFX96 Real-Time System (Bio-Rad) using a 30 sec denaturation step at 95°C, followed by 40 cycles of 5 sec at 95°C and 5 sec at 58°C. Finally, a melting curve was generated in 0.5°C increments for 5 sec from 65 to 95°C.

## **3.2 High-throughput genome-wide procedures**

### **Nuclei preparation and MNase digestion (MNase-seq)**

For each MNase digestion, 25 x 10<sup>6</sup> S2 cells were collected and washed with 10 ml of cold 1x PBS and centrifuged (500 g, 4°C, 7 min). The cell pellet was then resuspended in 1 ml of NP-40 lysis buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM MgCl<sub>2</sub>, 0.5% NP-40, 0.5 mM PMSF, 0.15 mM spermine, 0.5 mM spermidine) for nuclei extraction and permeabilization. The lysate

was incubated on ice for 5 min to complete the lysis and centrifuged (500 g, 4°C, 7 min). The pellet was washed once with MNase digestion buffer (10 mM Tris-HCl pH 7.4; 15 mM NaCl; 60 mM KCl; 0.5% NP-40, 0.5mM PMSF, 0.15 mM spermine, 0.5 mM spermidine) without resuspension and centrifuged again (500 g, 4°C, 5 min). The pellet was resuspended in 4.8 ml of MNase digestion buffer supplemented with CaCl<sub>2</sub> 1 mM and warmed up to 25°C for 5 min. From this resuspension, 800µl were used to test the digestion by using variation of the MNase digestion time. For each test, 100 µl of nuclei preparation were used with the same protocol described below for the bulk digest, but scaled down. For the bulk digest, 4 ml of nuclei preparation (stored on ice and at 4°C) were used and warmed up to 25°C for 5 min. 7.5 U of MNase (Sigma-Aldrich, N3755) were added to each digestion level and incubated for 1, 3 and 15 min for short-, typical- and long- (15') digestion levels. The reaction was stopped with 400 µl of stop buffer (1:1 of 0.5 M EDTA pH 8 and SDS 10%). Cross-linked samples were incubated at 65°C o.n. for de-cross-linking after this step. NaCl and sodium acetate at pH 5.2 were added to a final concentration of 400 mM and 300 mM, respectively. From this solution, DNA was isolated and purified using a commercial kit (QIAquick PCR purification kit, Qiagen) and eluted with 30 µl of 0.1x TE. RNase treatment was performed with 12 µl of RNase cocktail enzyme mix (Ambion, AM2288) at 37°C for 30 min. Di-, mono- and sub-nucleosomes were separated and cut out separately using a TAE 3% agarose gel. Agarose gel analyses conducted on prolonged digestion levels were performed on a TAE 4% agarose gel. The collected gel pieces were smashed mechanically, covered with the 'crush and soak' buffer (0.5 M ammonium acetate; 0.3 M sodium acetate pH 4.5; 0.1% SDS; 1 mM EDTA pH 8), and incubated at 37°C overnight. The solution was collected and the DNA was extracted and purified with the same Qiagen commercial kit.

Libraries were prepared with 200 ng of starting material by using reagents and protocol from NEBNext Ultra DNA Library Prep Kit for Illumina (NEB, E7370L). Index primers were purchased from the same company (NEB, E7335L). MNase-seq libraries were amplified with 7 PCR cycles and the DNA was purified during library preparation with Agencourt AMPure XP magnetic beads (BeckmanCoulter, A63881). Library concentration and fragment size distribution were assessed by Bioanalyzer High Sensitivity DNA kit (Agilent, 5067-4626).

### **MNase digestion of genomic naked DNA**

A scaled down protocol used for the MNase chromatin digestion was also applied to MNase digestion of genomic naked DNA (gDNA). gDNA from  $20 \times 10^6$  S2 cells was taken up in a volume of 650  $\mu$ l and digested with 2 U of MNase for 30'' or 3 min (gDNA short and long digestions). Each reagent was supplemented to have the final composition of the MNase digestion buffer. The digestion was stopped with 65  $\mu$ l of stop buffer. NaCl and sodium acetate at pH 5.2 were added to a final concentration of 400 mM and 300 mM, respectively. Downstream steps were identical to MNase digestion of the chromatin. The DNA was then separated in a TAE 3% agarose gel, size-selected at around 150 bp for gDNA short or totally extracted from the gel for gDNA long. Library preparation was performed with same parameters used for the MNase digested chromatin samples.

### **Sample preparation and shearing for X-ChIP-seq**

$50 \times 10^6$  cross-linked S2 cells were lysed in 1 ml of sonication buffer (20 mM Tris-HCl pH 8; 0.5% SDS, 2 mM EDTA; 0.5 mM EGTA; 0.5 mM PMSF; 1 tablet of proteases inhibitors, Roche 04 693 132 001) and incubated on ice for 10 min. The lysate was then transferred in milliTUBE 1 ml AFA fiber (Covaris, 520130) and sonicated with an E220 evolution machine (Covaris) with the following parameters (80 W / 20 df / 200 cpb / 2250 sec). The sonicated chromatin was centrifuged at (16100 g, 10 min, 4°C) and the supernatant was collected and quantified by Nanodrop. Shearing was checked by running the DNA in a 2% TAE agarose gel to evaluate the DNA fragment size distribution. Protocol was optimized in order to produce a smear centered at 200 bp with the majority of the fragments being less than 500bp. For rpb3 pull-downs, the epitope integrity was also checked by immunoblot.

### **Sample preparation and shearing for MNase-ChIP-seq**

Cells were cross-linked as indicated in paragraph 3.1, and nuclei preparation was performed through the same protocol described in the MNase-seq.  $25 \times 10^6$  cross-linked S2 cells were resuspended in 210  $\mu$ l of MNase digestion buffer. 2 U for 3 min and 7.5 U for 13.5 min were the MNase conditions used to shear the chromatin in short- and typical-digestion levels. The reaction was stopped with 10.5  $\mu$ l of stop buffer (20 mM EDTA; 0.1% SDS) and 1.1 ml of RIPA 600 buffer (20 mM Tris-HCl pH 8; 600 mM NaCl; 1% NP-40; 0.1% sodium deoxycholate; 0.1% SDS; 1 tablet of protease inhibitors, Roche 04 693 132 001) were immediately added. The digested lysate was kept on ice for 1hr. Samples were then centrifuged (15.000 g, 15min, 4°C) to

extract the soluble chromatin and 2.3x volumes of RIPA-no-salt buffer (20 mM Tris-HCl pH 8; 1% NP-40; 0.1% sodium deoxycholate; 0.1% SDS; 1 tablet of protease inhibitors, Roche 04 693 132 001) were added to the supernatant.

### **ChIP-seq of histone tail PTMs**

For each IP, 30 µg of chromatin (quantified by Nanodrop) was placed in 500 µl of RIPA (20 mM Tris-HCl pH8; 150 mM NaCl; 1% NP-40; 0.1% sodium deoxycholate; 0.1% SDS; 1 mM EDTA; 1 tablet of protease inhibitors, Roche 04 693 132 001). 10% of the chromatin used for the IP was collected as input, which DNA was de-crosslinked and purified following the same protocol used for the IP. The following antibodies were used for the immunoprecipitations: 2.5µg of Abcam Ab8580 (H3K4me3), Abcam Ab4441 (H3K9Ac), Abcam Ab1191 (H3K18Ac) and Abcam Ab4729 (H3K27Ac). IPs were performed in head-over-tail rotation at 4°C overnight. On the next day, 100 µl of dynabeads coated with protein G (Invitrogen 10009D) were prepared for each IP. Therefore, dynabeads were washed 3 times in RIPA buffer and incubated head-over-tail for 1hr at 4°C in 1ml of RIPA buffer plus 10 µl of Ultrapure BSA (50 mg/ml, Ambion AM2616) and 10 µl of Ultrapure Salmon Sperm DNA solutions (10 mg/ml, Invitrogen 15632011). Subsequently, beads were washed again 3 times for 5 min with RIPA and resuspended in a final volume of 100 µl of RIPA. Finally, beads were added to the IP and incubated for 3 hr to capture immuno-complexes. Subsequently the following washing steps were performed: 3 times for 3 min with RIPA; 1 time for 5 min with LiCl buffer (20 mM Tris-HCl pH 8; 250 mM LiCl; 1 mM EDTA; 0.5% NP-40; 0.5% sodium deoxycholate; 1 tablet of protease inhibitors, Roche 04 693 132 001); 1 time for 1 min with TE buffer; 1 rinse in TE in order to move the entire solution in a new DNA low binding tube. Successively, beads were resuspended in 100 µl of elution buffer (1% SDS; 0.1M NaHCO<sub>3</sub>), and de-crosslinked at 65°C for 2.5 hr at 1000 rpm. RNase treatment was then performed with 3 µl of RNase cocktail enzyme mix (Ambion, AM2288) at 37°C for 30 min at 1000 rpm. Following, a proteinase K treatment was performed with 60µg of the enzyme at 55°C for 1 hr at 1000 rpm. Finally, DNA purification and elution was performed by using commercial kit (MinElute PCR purification kit, Qiagen) following manufacturer's instructions. For each histone modification, 3 IPs were run in parallel and the immunoprecipitated DNA was the collect together during the elution step to increase the final yield. At this point, a quality control was carried out by qPCR to assess the enrichment of recovered fragments released from known promoter regions where these histone tail PTMs were found (modENCODE data) over known

inactive promoter (TSS of the *wnt10* gene). This evaluation was performed by calculating the fold change of the % input recovered from the IPs.

Libraries were prepared by using the entire eluted DNA derived from the 3 IPs, and from 5 ng of input DNA. Library preparation was performed using reagents and protocol from NEBNext Ultra DNA Library Prep Kit for Illumina (NEB, E7370L) and using index primers from the same company (NEB, E7335L). Libraries were cleaned using Agencourt AMPure XP magnetic beads (BeckmanCoulter, A63881) using a right side size-selection of 0.6x / 1.8x for the sonicated chromatin preparations, and of 0.75x / 1.8x for the MNase sheared chromatin preparations. Beads size-selections were performed after the PCR step. Libraries were amplified with 15 PCR cycles for the sonicated chromatin preparations and 13 PCR cycles for the MNase sheared chromatin preparations.

### **ChIP-seq of rpb3 (Pol II)**

ChIP of Rpb3 (a subunit of Pol II) was performed with minor changes from the sample preparation and ChIP protocols described above. Covaris sonication was performed with the following parameters (100 W / 20 df / 200 cpb / 1200 sec). 50µg of chromatin and 5 µl of homemade antibody (rabbit) were used for each IP. Mock IP was performed by using equivalent amount of normal rabbit IgG (Cell Signaling, #2729). The signal obtained from the mock IPs was not significant and with very few peaks along the genome. For such a reason, no further analyses regarding mock IPs were carried out. At this point, a quality control was performed by immunoblot and qPCR. For the former, input, eluate, beads, flow-throughs and washing steps protein precipitations were analyzed by SDS-PAGE electrophoresis using a 12% polyacrylamide gel and blotted against rpb3. qPCR assessment was performed to evaluate the enrichment of recovered fragments released from known promoter regions where Pol II was significantly enriched (TSS of *Actin5c*) over known inactive promoter (TSS of *CGI679I* gene). This was performed by calculating the fold change of the % input recovered in the IP.

Libraries were prepared from the entire eluted DNA derived by the 3 IPs and from 7.5 ng of input DNA. Library preparation was performed using the reagents and protocol from NEBNext Ultra DNA Library Prep Kit for Illumina (NEB, E7370L) and using index primers from the same company (NEB, E7335L). Libraries were cleaned using Agencourt AMPure XP magnetic beads (BeckmanCoulter, A63881) using a right side size-selection of 0.65x / 1.8x.

Beads size-selections were performed after the PCR step. Libraries were amplified with 15 PCR cycles.

### **DNase-seq**

DNase-seq protocol was performed as described previously (Vierstra et al., 2014), with minor modifications. This experiment is part of the PhD thesis of Andrea Ennio Storti and Marta Bozek from the Gaul lab. Briefly, nuclei were isolated from  $50 \times 10^6$  S2 cells pellets. Pellets were washed with 10 ml of cold 1x PBS and centrifuged (500 g, 4°C, 7 min). Then, pellets were resuspended in 2 ml of NP-40 lysis buffer, incubated 5 min on ice, and centrifuged again (500 g, 4°C, 7 min). The obtained nuclei pellets were resuspended in 5 ml of DNase buffer A and centrifuged (500 g, 4°C, 5 min). Supernatant was discarded and nuclei pellets were kept on ice until DNase I treatment. DNase I (Sigma-Aldrich, D4527) was diluted in 2.5 ml of DNase I digestion buffer to a final concentration of 25 U/ml. Nuclei were treated for 3 min at 37°C, and immediately after 2.5 ml of Stop buffer were added, followed by incubation at 55°C for 1 hour. Subsequently, 30 µl of RNase cocktail (Ambion, AM2286) were added and samples were incubated at 37°C for 1 hour. After controlling the proper digestion level on agarose gel, samples were loaded on top of a 10-40% sucrose gradient and centrifuged at high speed (34,000 rpm, 20°C, 24 hours) in a SW40Ti rotor (Beckman Coulter). Fractions from the gradient were recovered by a fractionation machine (500 µl per fraction). DNA fragments size in each fraction was assessed by agarose gel, and all the fractions containing DNA fragments <500 bp were pooled. Three volumes of QG buffer (Qiagen) and 1 volume of isopropanol were added, DNA purified on MinElute columns (Qiagen) and finally eluted in 24 µl of Elution buffer (Qiagen). At this point, a quality control was performed by qPCR, in order to assess the enrichment of recovered fragments released from known open regions (TSS of *Actin5c* and *αTub84B* loci) over known closed regions (3' UTR of *ed* and *ems* loci).

Library preparation for Illumina sequencing was performed with the NEBNext Ultra DNA Library Prep kit (NEB, E7370L) and using index primers from the same company (NEB, E7335L) according to manufacturer's instructions, starting with 150 ng of DNA. After the adapter ligation step, a size selection was performed with the AMPure XP beads (BeckmanCoulter, A63881) in order to enrich for fragments shorter than 150 bp. PCR amplification was carried out with 8 cycles. Final library purification was performed with AMPure XP beads.

**DTA-seq**

Any experimental procedures concerning DTA-seq on S2 cells were performed by Katja Frühauf as fundamental part of her PhD thesis. Therefore, I do not describe them here, but I refer to her thesis available in the faculty's archives.

**Next-generation sequencing**

All libraries were sequenced on an Illumina GenomeAnalyzer IIx in order to produce 50 bp pair-end reads. DNase-seq libraries resulted in 80 to 100 x 10<sup>6</sup> reads, whereas MNase-seq and ChIP-seq libraries resulted in about 40 x 10<sup>6</sup> reads. Next-generation sequencing was performed by the LAFUGA sequencing facility at the Gene Center LMU Munich.

### 3.3 Computational procedures

Computation procedures described in this paragraph were mostly performed by Mark Heron (Söding group) as part of a collaborative effort. Here, I report his description of such procedures as reference to his PhD thesis (available in the faculty's archive), where is possible to find further details. Moreover, some of the computational procedures I implemented myself.

**Mapping of MNase-seq data**

Reads were mapped with Bowtie2 (Langmead and Salzberg, 2012) (v2.1.0, parameters: -I 0 -X 1000 -p 4) to the flybase v5.53 *D.melanogaster* genome (Attrill et al., 2016). Regions where fragments cannot be uniquely mapped to (unmappable) were excluded from all analysis. These unmappable regions were identified by slicing the genome into overlapping 150-bp (roughly one nucleosome length) fragments, creating *in silico* paired-end reads from them, and mapping these against the genome. Excluding all reads that mapped to multiple genomic locations, the coverage was computed and only genome regions with a coverage of 150 (all generated fragments) were kept for the analysis. Extremely short or long (<30, >400) fragments were excluded from all analysis. For the analysis we tested, repeating them with a stricter *in silico* fragment-size selection did not affect the results. Generally, the nucleosome-dyad position is defined as the fragment center. In the case of di-nucleosome fragments, two dyad positions are estimated from the fragment ends. Each is placed half the average length of a mono-nucleosome fragment (same digestion level) from one end. The genome-wide coverage tracks were computed by extending

the dyad positions with  $\pm 73$  bps and summing the coverage per base pair. Such coverage tracks are an unscaled approximation of the nucleosome occupancy. The coverage tracks were normalized to a genome-wide average of 1, after doubling the counts on chromosome X, since S2 cells only have one copy due to their sex. Chromosome Y was removed completely, together with the heterochromosome regions of the other chromosomes, due to low mappability.

### **Nucleosome calling**

For the fragility and resistance analysis, nucleosome positions were called with the R package nucleR (Flores et al., 2014). Peaks were called independently on the mononucleosome fractions of the three digestion levels and merged in the final step. Fragments of length 50-200 bp were processed as described in the package vignette. Peaks were called for each dataset with a threshold of 25%. The peaks were filtered by their h-score which describes the height i.e. the amount of count data ( $>0.55$  for the group ‘all’ used throughout the main figures and  $>0.9$  for the group ‘only high’). The filtered peaks of the three datasets were joined and overlapping peaks were merged ( $<21$  bp between the peaks), using the joint center as the called nucleosome position. The filtering and merging was done to reduce error in the analyses.

### **Promoter calling**

15,971 promoters were assigned to 11,536 unique genes from clustering cap analysis of gene expression (CAGE) data (Brown et al., 2014). Transcript annotations were taken from the flybase v5.53 (Attrill et al., 2016). NP and BP promoters were distinguished by the dispersion of their transcription initiation. The promoters were classified into 8709 BP and 7262 NP promoters, based on the mean absolute deviation of the CAGE data mapped to the TSS. The promoter list was an updated version of the one used in Siebert and Söding (2016).

Processed RNA-Seq data was mapped from transcripts that started close ( $\pm 50$  bps) to the TSS and summed for each promoter. The promoters (BP and NP combined) were separated into quarters based on their assigned expression values. The 4th quarter was marginally extended to contain all unexpressed promoters.

### **Fragility and resistance score and nucleosome populations**

The fragility score was defined as the normalized mono-nucleosomal coverage divided by the normalized sub-nucleosomal coverage. Nucleosomes were assigned to the populations based on the fragility score of the typical digestion. Fragile: sub-nucleosome<sup>typical</sup> : mono-



nucleosome<sup>typical</sup> > 2; Resistant: mono-nucleosome<sup>typical</sup> : sub-nucleosome<sup>typical</sup> > 2. Average: nucleosomes neither fragile nor resistant. They also needed a minimum value of 0.5 for both coverages to reduce noise. Alternative scores were measured with same procedure using mono-short- : mono- long- (15') (monos) and dinucl- short- : mono- short- (oligos) ratios, as explained in paragraph 4.3.8.

### **Computing gene distances**

Gene organizations were calculated by an updated version of the promoter list used in Siebert and Söding (2016) complemented with TSS and TTS coordinates extracted from flybase v5.53 *D.melanogaster* genome (Attrill et al., 2016). In this analysis, alternative promoters were excluded by considering only the promoter with higher CAGE counts per gene. In these lists, promoter distinction and expression quartile was kept as described above. Composite plots of gene distance clusters plus the ones shown in Part III were calculated and visualized with DeepTools (Ramirez et al., 2014), using the “computeMatrix” and plotProfile” tools.

### **Mapping and peak calling for DNase-seq**

Mapping and peak calling was performed by Roberto Cortini, a PhD student in our lab. Briefly, after de-multiplexing, 50 bp reads were trimmed to 27 bp from the 3' end to remove adaptors and maximize alignability. To map trimmed reads, Bowtie2 (Langmead and Salzberg, 2012) was used with the following parameters: --local --very-sensitive-local --threads 16 --maxins 1000; default parameters if not specified. Release 5.53 of the *D.melanogaster* reference genome was used. Alignment results were filtered for a minimum MAPQ score of 13 and for proper pairing using SAMtools (Li et al., 2009): samtools view -f 0x3 -q 13. Peak calling was done using MACS2 (Zhang et al., 2008) with the following parameters: callpeak --keep-dup all --nomodel --shift -100 --extsize 200 -f BAM -g dm and using a naked DNA sample as control file.

## 4 RESULTS

### 4.1 Study design

#### 4.1.1 Experimental design

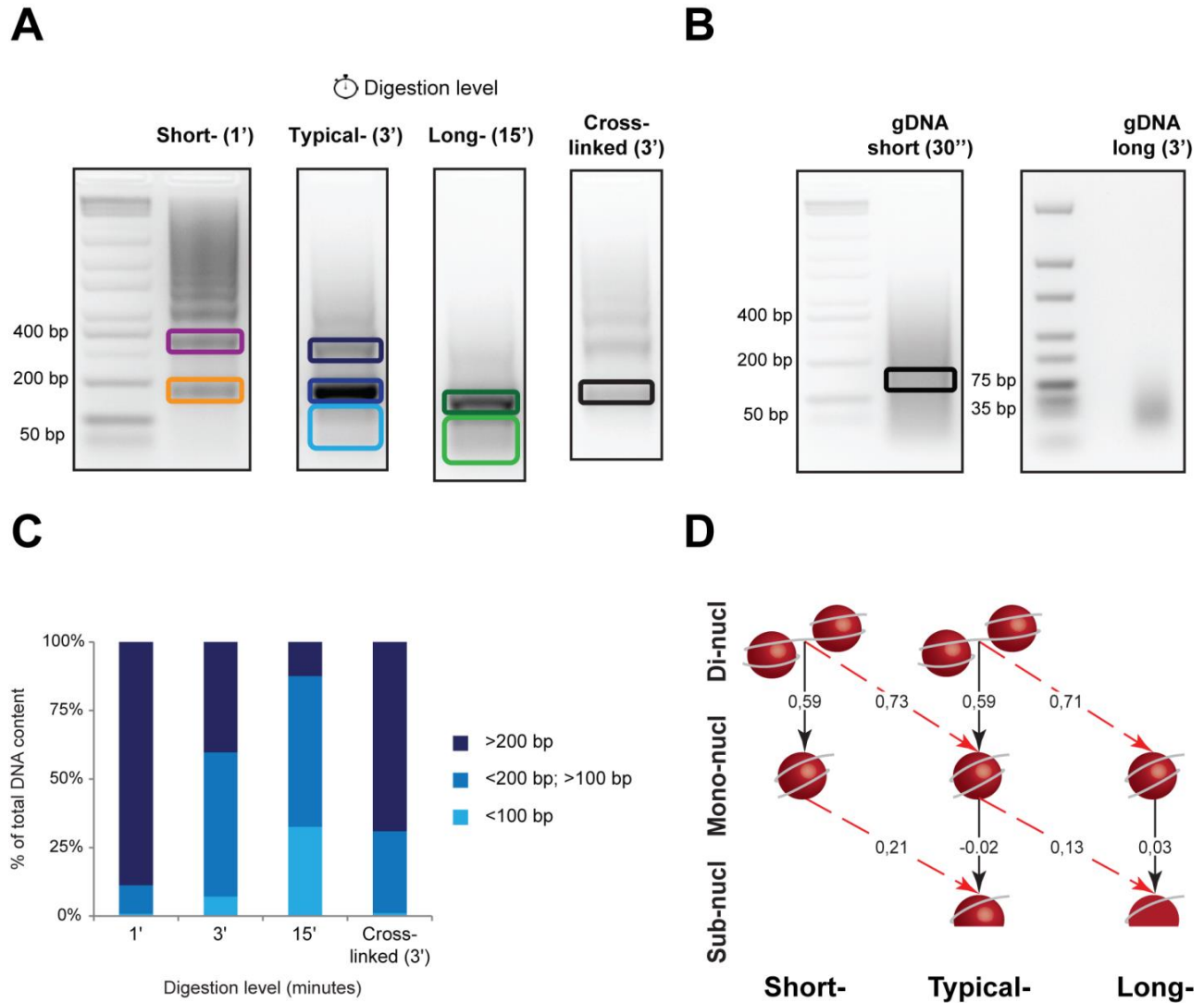
A fundamental step to study the MNase sequence bias is the investigation of the enzymatic activity on genomic DNA. MNase-digested genomic DNA provides powerful information on cutting events occurring in a chromatin-free environment. Thus, modes of action of the enzyme driven by the chromatin structure are eliminated. Genomic DNA was purified through a high-salt preparation (Aljanabi and Martinez, 1997) and two digestion levels were applied. **gDNA short** (30''), which produces a smeared DNA fragment distribution centered at the nucleosomal DNA size. Such a DNA size was isolated and sequenced, providing a greater control for the mono-nucleosomal map. **gDNA long** (3') was obtained using the same digestion time of a typical MNase digestion of the chromatin without any size-selection, which provides insights on cuts within disfavored sequences. The gDNA long fragment distribution is shorter than 75 bp and is centered at around 35 bp (Figure 3.B).

In parallel, to study MNase digestion inside the nucleosomes, prolonged digestions were carried out, leading to particles of the size of mono-nucleosomes or shorter. The digestion was performed with both native (9' and 30') and FA cross-linked (15' and 45') chromatin. Mono- and sub-nucleosome were isolated and sequenced without any size selection (Figure S1.B). FA cross-linking presumably strengthens DNA-histone interactions, thereby providing an additional condition to study MNase digestion on the chromatin. The digestion time is different between native and cross-linked chromatin to match digestion levels. Experimental comparison between native and cross-linked preparations was performed through quantification of the DNA signal from the agarose gel by matching nucleosomal fractions.

In this study, differential MNase-seq was performed with three digestion levels: **short-**, **typical-** and **long-**digestion. From each digestion level, **dinucl-**, **mono-** and **sub-**nucleosome fractions were isolated and sequenced separately in order to have similar genome coverage

(Figure 3.A). The ratio between fractions in each digestion level was quantified by agarose gel signal (Figure 3.C). This approach to evaluate digestion level is more reliable than using only the digestion condition, due to differences in efficiency among enzyme batches. Relative ratios were also confirmed by Bioanalyzer quantification. Short-digestion was obtained with 1' of MNase digestion and produced mostly oligo-nucleosomes, small amount of mono-nucleosomes and insignificant amount of sub-nucleosomes (about 89%, 10% and 1%, respectively). Typical-digestion was obtained with 3' of MNase incubation and produced mostly oligo- and mono-nucleosomes, with a small amount of sub-nucleosomes (about 40%, 53% and 7%, respectively). Long-digestion was obtained with 15' of MNase incubation and produced small amount of oligo-nucleosomes, but a higher amount of mono- and sub-nucleosomes (about 12%, 55% and 33%, respectively). Fragment distribution of **dinucl- short-** was characterized by a sharp peak around 350-360 bp, which is compatible with a protection of two chromosome joined by a linker DNA (Figure S1.A). Instead, fragment distribution of **dinucl- typical-** was characterized by a peak around 310-320 bp, most likely due to an exhaustive digestion of the flanking linker DNAs and/or digestion at the entry/exit of one or both nucleosomal DNAs. A similar scenario was observed for the mono-nucleosome fraction. In that case, the fragment distribution was characterized by a peak around 165, 150 and 145 bp for **mono- short-**, **mono- typical-** and **mono- long-**, respectively. The mono- short- length mirrors a chromosome, whereas the fragment distribution of mono- long- can derive from a partial digestion at the entry/exit of the nucleosomal DNA. In parallel, a more complex pattern in the fragments distribution of **sub-typical** and **sub- long-** was observed, although still characterized by a reduced averaged fragment length with the increase of the digestion level. Furthermore, to measure the influence of FA treatment on MNase digestion and nucleosome mapping, cross-linked chromatin was digested with the typical- digestion time. Its digestion level was between the short- and typical-digestion levels of native chromatin, with a proportion among oligo-, mono- and sub-nucleosome fractions of 69%, 30% and 1%, respectively. For the FA sample, only the mono-nucleosome fraction was sequenced.

# Figure 3



**Figure 3: Differential MNase-seq design.** (A) Fragment length separation on TAE agarose gel 3% of three MNase digestion levels on the chromatin. Marked boxes indicate samples that were extracted, sequenced and analyzed. (B) Fragment length separation on TAE agarose gel 3% of MNase digested genomic naked DNA (gDNA). In gDNA short (30''), the black box showed the size-selection used for this sample. (C) Quantitative evaluation of the DNA signal from (A) performed with ImageJ. Fragment length separation was obtained through comparison with the ladder. Fragments longer than 200 bp contain oligo-nucleosomes, between 200 and 100 bp contain mono-nucleosomes and below 100 bp contain sub-nucleosomes. (D) Selected pairwise Pearson correlations between genome-wide coverage of some MNase digested native chromatin samples. The arrangement matches with (A). Black unbroken arrows indicate correlation values among fractions of the same digestion level; red dashed arrows indicate correlation values among fractions of different digestion levels.

### 4.1.2 CREs definitions

As mentioned in the introduction, compelling evidences show that differential MNase-seq is a powerful tool to characterize chromatin features in regulatory regions. Therefore, a similar strategy was pursued around active and inactive CREs. As chromatin landscape varies according to the state and feature of each CRE, a complex classification was performed based on activity, shape and genic organization for promoters.

Promoters were divided in three main groups according to promoter shape and genic organization: **NP**; **BP directional**; **BP divergent** (Figure 2.A). BP promoters, in which TSSs are dispersed over tens of bps, are typically found in constitutively expressed genes and have a canonical nucleosome pattern (Rach et al., 2011). NP promoters, in which TSSs are sharply defined within a few bp, are typically found in inducible genes with high expression plasticity and their nucleosome patterns are non-canonical (Rach et al., 2011). Promoter identification and NP/BP clustering was performed using CAGE data (Brown et al., 2014). 15,971 promoters were assigned to 11,536 unique genes and classified into 8709 BP and 7262 NP promoter shapes. BP promoters were further divided in two clusters based on the gene organization. BP divergent cluster included protein coding genes transcribed in opposite direction and with a distance less than 500 bp between the two TSSs. The remaining BP genes were clustered as directional. Moreover, genes were divided in quartile ascending order according to their expression level, thus the first quartile contains highly expressed genes (Figure S1.C). Expression data were produced through DTA-seq on total RNA in S2 wild type cells.

As enhancers, **DHS** peaks at least 500 bp away from any TSS and TTS were considered in order to include only distal regulatory elements (Figure 2.B). DHS peaks were divided in quartile of signal strength (Figure S1.C). Moreover, DHSs were further classified in three classes according to their state: enhancers that are completely shut down in S2 cells, but active in ovary stem cells (OSC) (**OSC unique**) (Arnold et al., 2013), active enhancers in both the two cells lines (**OSC/S2 shared**), and uniquely active in S2 cells (**S2 unique**).

Finally, **TTSs** were divided according to the promoter shape of the gene: BP and NP. Once again, only TTSs at least 500 bp away from any TSS were considered.

## 4.2 PART I – MNase biases

In the first part of this study, the two main biases that affect MNase digestion were investigated.

Firstly, MNase has a strong sequence bias with an A or T at the 5' end at the cut site. However, earlier work demonstrated that the cut site is not simply influenced by the AT content, but rather by a broader sequence context (Dingwall et al., 1981; Horz and Altenburger, 1981). Nonetheless, these studies were performed on a limited set of sequence, thereby a detailed sequence bias characterization at the cut site is required genome-wide. Moreover, to map the sequence bias along the genome, a MNase digestion on genomic naked DNA (gDNA) was used as control, since it assesses the preference of cutting along the genome without protein protection (Chung et al., 2010). The gDNA control was mostly used in yeast and analyses were carried out mostly by genome-wide correlation with MNase digested chromatin samples, or by track overlapping on specific loci. Therefore, little information is available on the sequence bias along the *D.melanogaster* genome, and a specific characterization of it around CREs is also missing.

Secondly, this part of the thesis also investigated how MNase cuts within the nucleosome. In this case, two questions were asked: 1) MNase cuts preferentially by a continuous or discrete digestion; 2) MNase cuts preferentially on bound or unbound (unwrapped) DNA. Therefore, MNase was used to probe features of DNA-histone interactions by assessing nucleosome accessibility genome-wide.

### 4.2.1 MNase sequence bias at the cut site is complex

As mentioned above, a characterization of the sequence bias at the cut site was firstly performed on gDNA by using two digestion levels: gDNA short and gDNA long. This evaluation was assessed by both PWM analysis and di-nucleotide frequencies calculation at the cut site.

In gDNA short, a position weight matrix (PWM) surrounding the cut site was extracted, with the consensus sequence aTAg and an information content of 1.30 (Figure 4.A). In gDNA long, the extracted PWM was weaker, with the consensus sequence atAg and lower information content (0.67); consistent with the notion that MNase initially cuts favored sequences, followed by less favored ones. For such a reason, gDNA short better indicates MNase sequence preference of cutting due to the availability of a more variegate sequence pool. To obtain more insights into the preferred local sequence environment, fold changes of all 16 di-nucleotides against their

genome-wide average frequency were calculated (Figure 4.B). All di-nucleotide frequency plots were produced through fragment alignment by their start at the 5', thereby the +1 position corresponds to the first sequenced nucleotide. At the cut site, a strong enrichment for TA, a moderate enrichment for AA and CA and a strong depletion for SS di-nucleotides were observed in -1 position. These findings indicate a strong preference of cutting within a TA with a minor contribution played by the other WW di-nucleotides in the immediate surrounding (position -2 and +1). The preferred flanking sequences are different between upstream and downstream of the cut site. In the upstream context, enrichment for TA and moderate enrichment for AT, CT, TC and CC di-nucleotides were observed. We hypothesize these di-nucleotides as alternative sites for a first cut before the pseudo-exonuclease digestion, or alternatively as a sequence platform that enhances the probability of cutting in a downstream TA. Surprisingly, in the same context, no enrichment and even moderate depletion for AA and TT di-nucleotides were measured. In the downstream context, a more homogenous situation was observed, with moderate depletion in WW and enrichment in SS di-nucleotides. Interestingly, at the position +1 a strong enrichment for AG was obtained, which is followed by an almost exclusive enrichment for GN di-nucleotides at the position +2 and +3. Additionally, a moderate depletion for CN di-nucleotides was also obtained in the same positions. These results indicate a potential and exclusive role for the G nucleotide as a blocker of the MNase pseudo-exonuclease activity on the forward strand. The sequence cut site landscape is less heterogeneous in longer digestions. Indeed, gDNA long showed a more prominent dichotomy between WW and SS di-nucleotides into the fragments, even if most of the enrichments or depletions measured in gDNA short are preserved. This could be explained by an accumulation effect of the sequence bias.

In conclusion, as already observed in low-throughput studies, the sequence complexity around MNase cut sites is also confirmed genome-wide. Therefore, MNase sequence bias is not only driven by the AT content, but rather it shows a strong dependency for the TA di-nucleotide, although surrounded by a complex and non-homopolymeric landscape.

#### **4.2.2 MNase sequence bias in chromatin is reduced**

Given the complexity of the MNase sequence bias at the cut site of the gDNA, similar analyses were performed on MNase digested chromatin samples to ascertain changes in the MNase sequence bias.

## *Results*

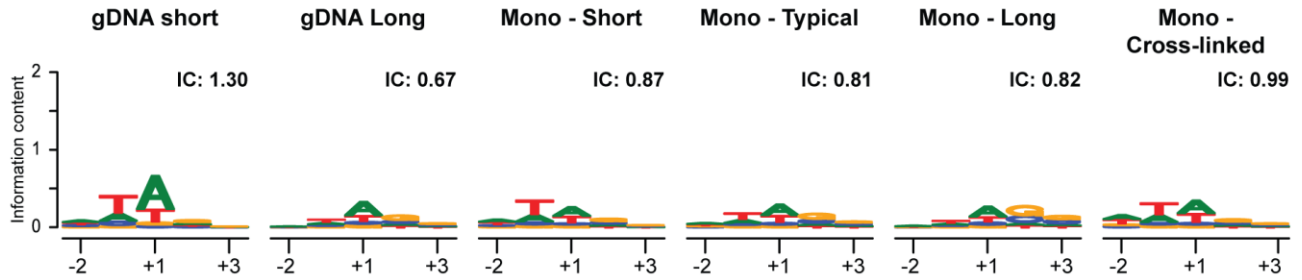
In chromatin-digested samples, the sequence signature aTAg was still visible, but the relative base contributions changed and lower information contents than gDNA short were measured, decreasing as a function of fraction size and digestion time (Figure 4.A & Figure S2.A). However, when chromatin is shortly digested, the nucleosomal content partly follows the sequence bias. This derives by a higher availability of preferred cutting sites, and from an accumulation effect that reduces the heterogeneity of the sequence context around them. Therefore, a reduction in TA frequency at position -1 for each fraction was detected with the increase of the digestion level (Figure 4.B & Figure S2.B). Moreover, a decrease in complexity around the cut site was also observed with longer digestion levels, which can be mostly explained by the WW and SS dichotomy. In particular, a more evident role of the G nucleotide as blocker of the MNase trimming emerged, indicated by the increase of the G contribution at the position +2. Similar results were obtained for both di- and sub-nucleosomal fractions. Sub-nucleosomes were characterized by a more prominent dichotomy between WW and SS di-nucleotides, mostly upstream of the cut site. Indeed, sub-nucleosomes are mostly generated by digestion inside the nucleosome. Consequently, their di-nucleotide frequencies partially mirror the sequence context of the nucleosomal DNA from which they derived.

Taken together, MNase sequence bias at the cut site on the chromatin conserve most of the features observed in gDNA, but it is reduced, most likely due to the occlusion of most of the genomic DNA by nucleosomes.

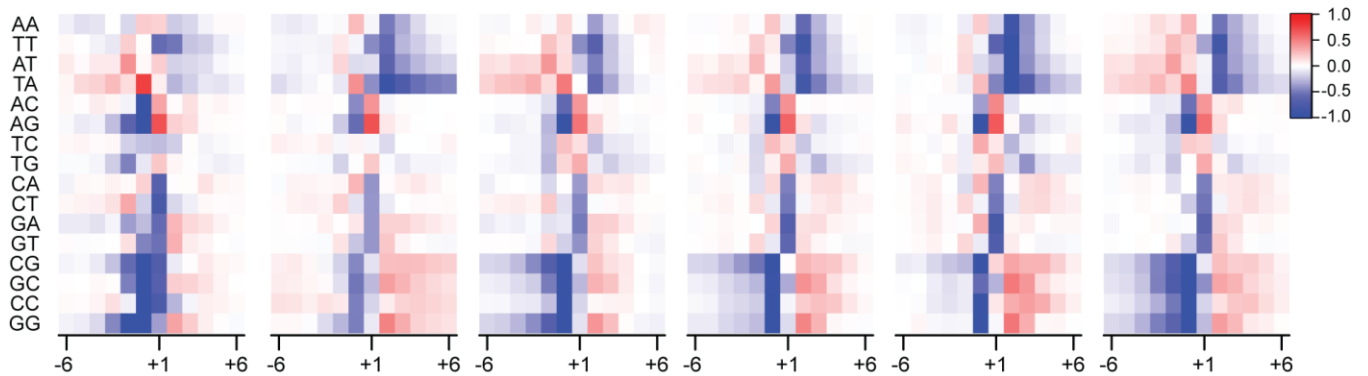


# Figure 4

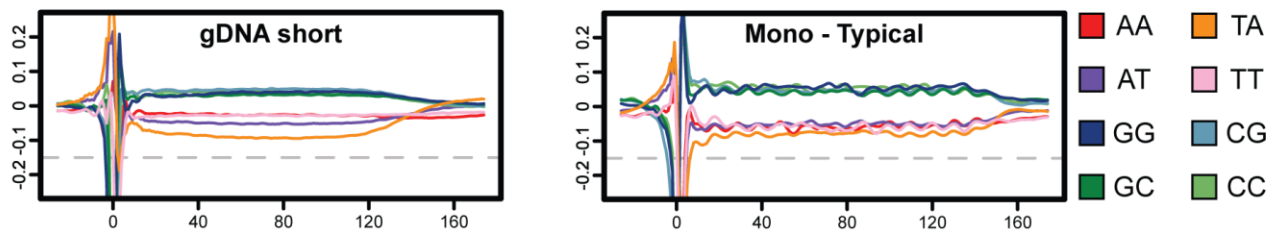
**A**



**B**



**C**



**Figure 4: MNase sequence bias at the cut site.** (A) PWMs around MNase cut sites of genomic DNA (gDNA) and mono-nucleosomal samples. The position +1 corresponds to the start of the fragments. The sum of the depicted PWMs information content (IC) is shown in the right corner of each individual panel. (B) Di-nucleotide enrichments around the cut sites. The color scale shows the log fold change against the genome-wide average frequency. PWMs and di-nucleotide enrichments around the cuts site for other chromatin samples are reported in Figure S2 and Figure 6. (C) Smoothed di-nucleotide enrichments over the fragment region aligned by the left MNase cut site – position 0. The left panel shows the gDNA control, which had enrichment for G and C (SS) di-nucleotides, but no 10 bp periodicity. The right panel shows mono- typical-. It is characterized by both SS enrichment and 10 bp periodicity described for nucleosomes.

### 4.2.3 MNase sequence bias does not determine nucleosome mapping around promoters

In the next step, MNase sequence bias was measured into the sequenced fragments and around CREs. For the former, the TA di-nucleotide was mostly considered, which MNase preference was demonstrated in the previous two paragraphs. In contrast to genomic DNA, nucleosomal DNA fragments show a pronounced 10 bp periodicity in their di-nucleotide frequencies (Figure 4.C). This indicates both the quality of the gDNA subjected to MNase digestion, in terms of proteins depletion, and the nucleosomal origin of the chromatin-digested fragments.

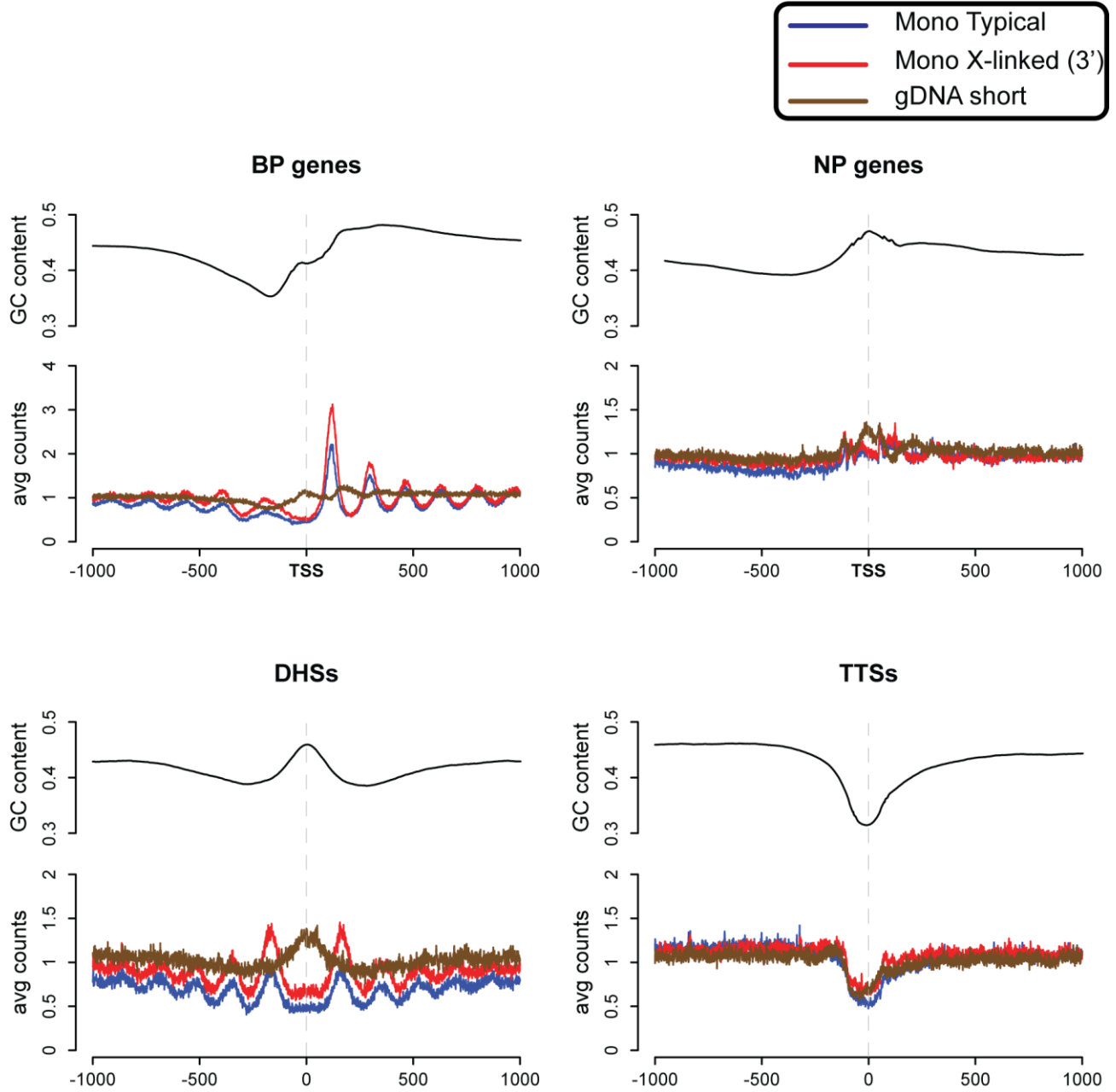
MNase digestion of the chromatin initially proceeds by cleaving protein-free DNA regions, such as linker DNAs. In this regard, a significant increase in TA di-nucleotides between uncut and cut linker DNAs was observed, both for dinucl- short- and dinucl- typical- (Figure 14.B). The same observation can be made for the mono-nucleosomal fraction, where the TA frequency in the linker DNA, downstream of the position 150, decreased from short- to typical- to long-digestion levels (Figure 14.C). Finally, in the di-nucleosomal fraction, the frequency of TA in both uncut and cut linker DNAs decreased with the increase of the digestion level, most likely due to an extinguishing of the TA availability. These observations indicate that nucleosomes surrounded by linker DNAs rich in TA are more easily isolated from the chromatin during the early steps of MNase digestion, potentially biasing the nucleosomal content in the mono-nucleosome fraction among digestion levels.

As noted in a previous study (Chung et al., 2010), the short digestion of genomic DNA with a size-selection of ~150 bp correlates with the mono-nucleosomal map and GC content, both at the single locus and genome-wide levels. Indeed, Pearson correlations (PCC) between gDNA short and native chromatin fractions were moderately high for mono-nucleosome (around 0.6), lower for di-nucleosome (around 0.52) and not significant for sub-nucleosome tracks (0.14 and 0.22 in typical- and long-digestion, respectively) (Figure S2.C). Notably, the PCC between GC content and gDNA short was modest (0.4). Nonetheless, it was higher compared to mono-short-, and lower compared to mono- of typical- and long-digestion levels (0.29, 0.58 and 0.71 respectively). Prompted by these results, gDNA and mono- typical- tracks were compared around BP and NP promoters (Figure 5). Interestingly, the two tracks were phase-shifted against each other in BP promoters. The nucleosomal DNA showed depletion within the promoter, whereas the genomic DNA had a peak. Furthermore, the -1 nucleosome was characterized by the lowest signal in gDNA, and the array downstream of the TSS showed lower amplitude and anti-

correlation compared to the nucleosomal track. A similar outcome was measured around NP promoters, although a comparison is weaker due to a reduced nucleosome phasing obtained from composite plot. As a final step, the same comparison around other CREs, such as DHSs and TTSs was also considered. In those regions, a higher overlapping between gDNA and nucleosomal tracks was found. Nevertheless, chromatin structures typical of DHSs and TTSs, included the NDR in the latter, were demonstrated as such through chemical mapping, which is not affected by the sequence bias (Brogaard et al., 2012). In this regard, it is likely that at some degrees the overlapping within these regions can be coincidental and driven by the correlation between GC content and nucleosome occupancy.

In conclusion, genome-wide correlation between gDNA and nucleosomal tracks can mask important details that emerge when the two tracks are compared on specific genomic regions. Indeed, our data reject the idea that nucleosome mapping is solely driven by MNase sequence bias, in particular around promoters.

# Figure 5



**Figure 5: Comparison between gDNA and mono-nucleosome from native and cross-linked chromatin.** Smoothed profiles of GC content (first line of each panel) and average reads counts (second line of each panel) for gDNA and mono-nucleosome tracks from native (typical-digestion level) and cross-linked chromatin (3') around BP promoters (top left), NP promoter (top right), DHS peaks (bottom left) and TTSs (bottom right).

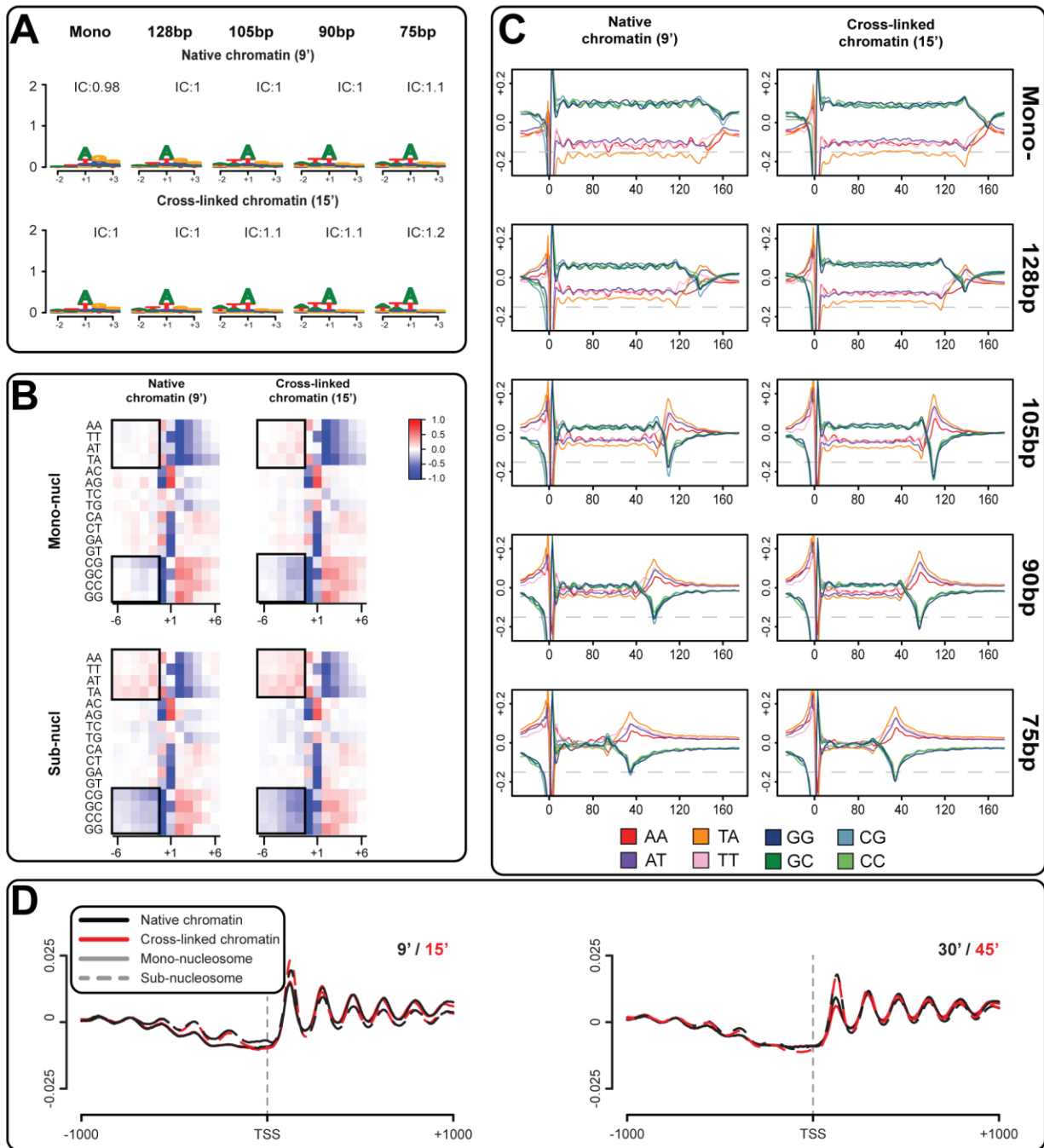
#### 4.2.4 MNase digestion on FA cross-linked chromatin

FA cross-linking was used to measure how MNase cuts on the nucleosomal DNA after DNA-histone interactions strengthening. Hence, MNase sequence bias differed on cross-linked chromatin was characterized to assess eventual differences with what was observed on native chromatin.

MNase sequence bias in cross-linked chromatin showed a more pronounced WW and SS dichotomy around the cut site. In fact, the highest PWM information content among mono-nucleosome samples was measured on cross-linked chromatin 3' (0.99). In the upstream context, a higher A and T contribution and a reduced complexity were also detected (Figure 4.A). In order to remove the digestion level difference between samples in this analysis, mono- and sub-nucleosome fractions were compared from similar digestion levels between native and cross-linked chromatin. The PWMs at the cut site were very similar, with only little changes in the information content. Nevertheless, an increased WW/SS dichotomy in the upstream context was detected in the cross-linked chromatin (Figure 6.A & Figure 6.B). This suggests a reduction of the MNase pseudo-exonuclease activity when cross-linked chromatin is digested, leading to cutting sites that more resemble features observed in gDNA short. Interestingly, a less pronounced 10 bp WW and SS 10 bp (anti) periodicity was also measured in cross-linked mono- and sub-nucleosome fragments (Figure 6.C). It could be possible that, in native chromatin, nucleosomes can re-adjust and find their most favorable rotational position during sample preparation and subsequent MNase digestion steps. Alternatively, a poor fragments alignment can also explain this observation, since differences are subtle. gDNA short genome-wide correlations with either native or cross-linked chromatin were not significantly different (Figure S2.C). Finally, no differences were detected both in nucleosome positioning and occupancy between native and cross-linked chromatin around CREs, when similar digestion levels were compared (Figure 6.D).

Collectively, MNase digestion on cross-linked chromatin shows a less complex sequence bias around the cut site with a more pronounced contribution played by A and T di-nucleotides. At the same time, FA cross-linking does not introduce differences in nucleosome mapping.

## Figure 6



**Figure 6: Comparison of MNase digestion from native and cross-linked chromatin.** (A); (B) Subfigures matching those in Figure 4.A and Figure 4.B, but for mono- and sub-nucleosome peaks from prolonged digestion levels of native and cross-linked chromatin. (C) Smoothed di-nucleotide enrichments over the nucleosome region for mono- and sub-nucleosome peaks from prolonged MNase digestion of native and cross-linked chromatin (see Figure 4.C). (D) Smoothed and normalized nucleosome dyad frequency calculated from fragment occurrences of mono- (unbroken lines) and sub-nucleosome (dashed lines) fractions around BP promoters. In each panel matched digestion levels between native (black lines) and cross-linked (red lines) chromatin preparations are shown.

#### 4.2.5 MNase digestion within the nucleosome is mostly discrete

Next, MNase digestion within the nucleosome was characterized by using prolonged digestion levels. The sub-nucleosome generation is a relevant step of the MNase flowchart, since it could lead to information loss when a mono-nucleosome size selection is applied or when fragments are too short to be analyzed. Additionally, it could lead to data misinterpretation, since shorter fragment could derive from non-histone proteins protection.

Interestingly, the sub-nucleosome fragment distribution from the Bioanalyzer showed three sharp peaks centered at 128, 105 and 90 bp and one broader at 75 bp (Figure 7.A). The distribution was also consistent with the length distribution of sequenced fragments from prolonged digestion levels (Figure S1.B) and with the DNA signal from the agarose gel (Figure 8.A). Analyses at single sub-nucleosomal peaks were performed through an *in silico* size-selection shown in Figure S1.B. Four observations suggest that most of sub-nucleosome fragments, at least longer than 75 bp, derive from nucleosome protection: 1) their formation occurred with a digestion level (around 3') comparable to gDNA long. The fragment distribution of the latter was shorter than 75 bp, with a peak at around 35 bp, indicating that protein free DNA is quickly digested away (Figure 3.B); 2) Non-histone protein protection from MNase was characterized by a fragment distribution shorter than 75-50 bp in several studies, mostly conducted by the Henikoff lab. Such fragments are mostly not included in this study; 3) di-nucleotide frequencies of sub-nucleosome fragments showed a nucleosomal 10 bp periodicity not only within the fragments, but also in the surrounding regions that were digested away (Figure 6.C); 4) Sub-nucleosomes pulled-down by ChIP-seq of histone tail PTMs were characterized by similar di-nucleotide profile (Figure S9).

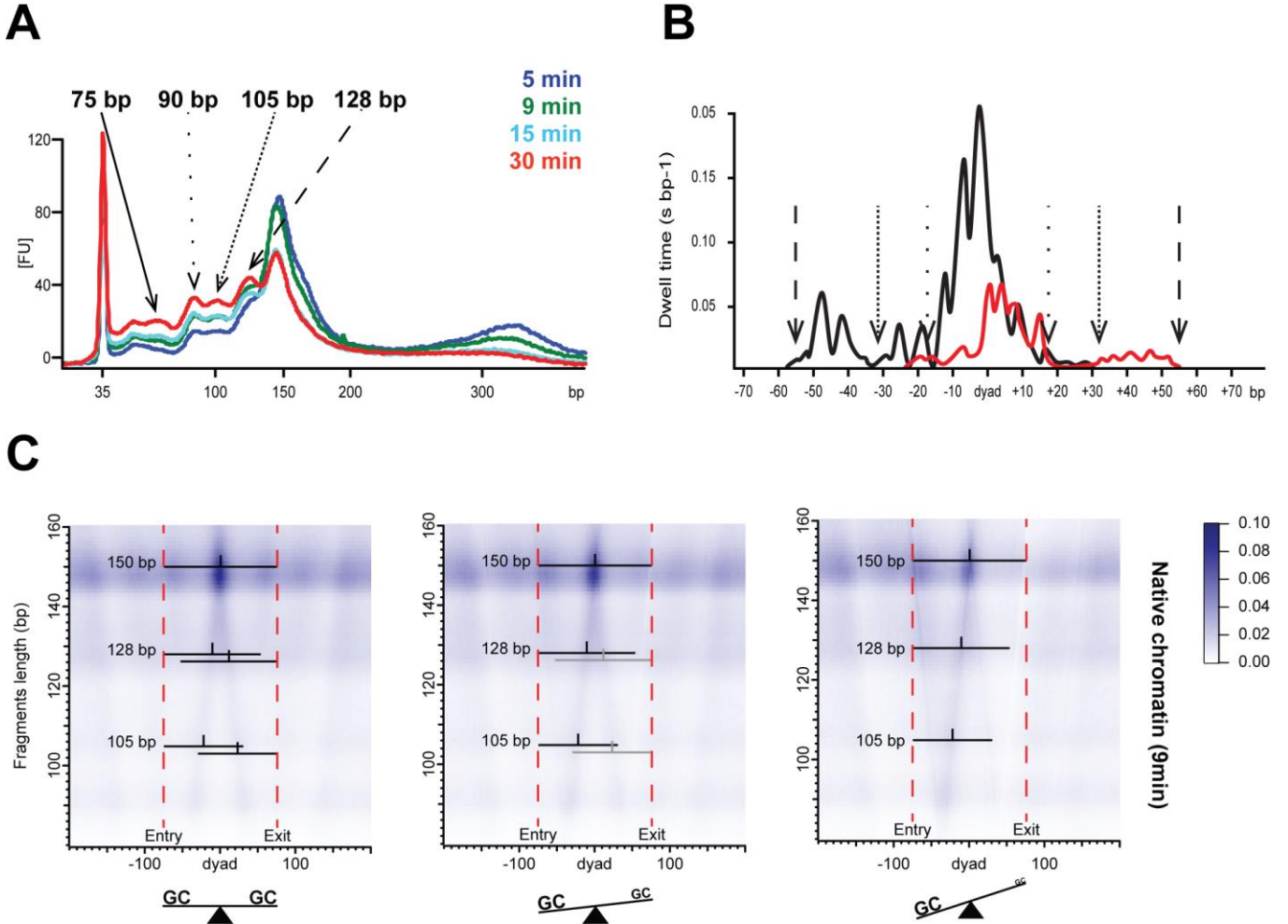
Lengths of the sub-nucleosome distribution peaks were consistent and reproducible within a large range of digestion levels, even after a MNase incubation 20 fold higher than typical digestion level (Figure 8.A). The peak at 128 bp was consistently the highest in intensity, most likely due to its origin from a cut closer to the nucleosome entry/exit. In parallel, peak intensities at 105 and 90 bp were comparable. Interestingly, the ratio between these peaks was also roughly preserved among digestion levels, suggesting a common source. Indeed, it is likely that most of sub-nucleosome fragments derive from the mono-nucleosome fraction, whose peak constantly decreased throughout digestion levels. To further confirm the consistency of the sub-nucleosome distribution, a differential MNase-seq based on concentration titration was next

performed. Also with this setup, both peaks length and ratios among them were preserved (Figure 8.B). Finally, variations in the sub-nucleosome fragment distribution were measured after DNA-histone interactions strengthening by FA cross-linking (Figure 8.C). Although the peaks length was preserved, the ratios among them changed, favoring the protection of longer sub-nucleosomes (128 and 105 bp) and disfavoring shorter ones (90 and 75 bp). Therefore, a higher protection against MNase toward the nucleosome dyad is proposed, in which the DNA-histone interactions are stronger and consequently more prone to form FA bridges.

In conclusion, MNase digestion within the nucleosome is not only a continuous process, but also proceeds in a discrete way, producing a fragment length distribution consistently characterized by favored lengths. Therefore, intra-nucleosome digestion seems to target certain nucleosome coordinates, most likely driven by variations in DNA-histone interactions strength.

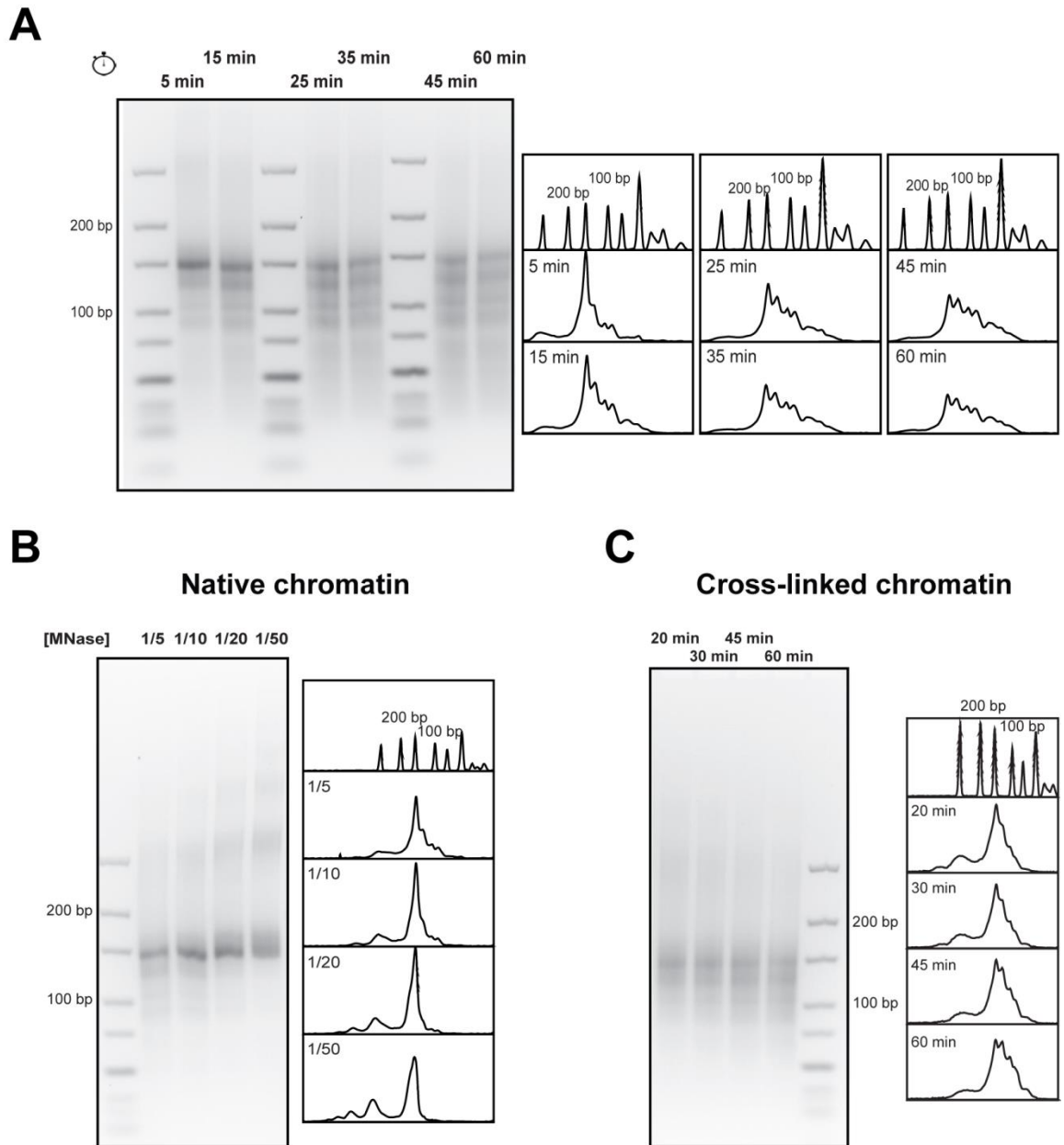


# Figure 7



**Figure 7: MNase digestion within the nucleosome.** (A) Bioanalyzer profiles of the bulk MNase digestion of native chromatin with different digestion levels. Fragment length distribution of sub-nucleosomes shows a presence of four peaks emerging from a continuous distribution. Peak sizes are confirmed by agarose gel analysis and length distribution of sequenced fragments. (B) Reproduction of a figure by (Hall et al., 2009). It shows profiles of the dwell time in mechanical unzipping from both sides of reconstituted nucleosomes with the 601-Widom sequence. Arrows with distinct dashed pattern match the fragment length showed in (A) originated by asymmetric MNase digestion. (C) V-plots showing the fragment-center frequency ordered by fragment length around called nucleosomes by nucleR (threshold = 0.55). For an even GC content ratio between nucleosome halves (left panel) MNase digestion within the nucleosome is asymmetric, but both halves are stochastically involved. This leads to an inverted V pattern. Asymmetric MNase digestion within the nucleosome involves mostly the half with a reduced GC content (middle panel) and this effect is more pronounced with the increasing of the GC content asymmetry between the two nucleosome halves (right panel).

# Figure 8



**Figure 8: Sub-nucleosome fragment distribution.** (A) Prolonged MNase digestion levels were performed with different digestion time on native chromatin and run in bulk in a TAE 4% agarose gel. (B) Differential MNase-seq was performed through titration of MNase concentration and run in bulk in a TAE 4% agarose gel. The shown numerical fractions are related to the MNase concentration used in our standard differential MNase-seq protocol. (C) Prolonged MNase digestion levels were performed on cross-linked chromatin and run in bulk in a TAE 4% agarose gel. On the right of each panel it is shown sub-nucleosome fragment distributions through analysis of the DNA signal with ImageJ after inverting the picture and removing the background. Fragment lengths are evaluated by DNA ladder comparison.

#### **4.2.6 MNase digestion within the nucleosome is mostly asymmetric**

MNase digestion within the nucleosome could happen while DNA is bound to the histones and/or following temporary accessibility driven by thermodynamic nucleosome motions, such as spontaneous unwrapping, or ATP-consuming active mechanisms. Although none of those hypotheses can be fully rejected, our results indicate a major role played by spontaneous unwrapping.

A recent biophysical study showed that nucleosome unwrapping proceeds mostly asymmetrically (Ngo et al., 2015). To test this notion at genome-wide fashion, it was investigated how sub-nucleosomes were distributed within nucleosomes. Sub-nucleosomal fragment centers were plotted around called nucleosomes and ordered by their fragment length (V-plot) (Henikoff et al., 2011) (Figure 7.C). The sub-nucleosome fragment length distribution was retained, since a condensation of fragment centers emerged around the length of 128, 105 and 90 bp. Most importantly, sub-nucleosomal centers bifurcated with the decrease in length: one in which the entry site remained intact and the exit site was digested, and the other in which the exit site remained intact and the entry site was digested. The asymmetric digestion was even accentuated both with the increase of the MNase digestion level and when cross-linked chromatin was used (Figure S3), probably due to an accumulation and preservation of sub-nucleosomal fragments, respectively.

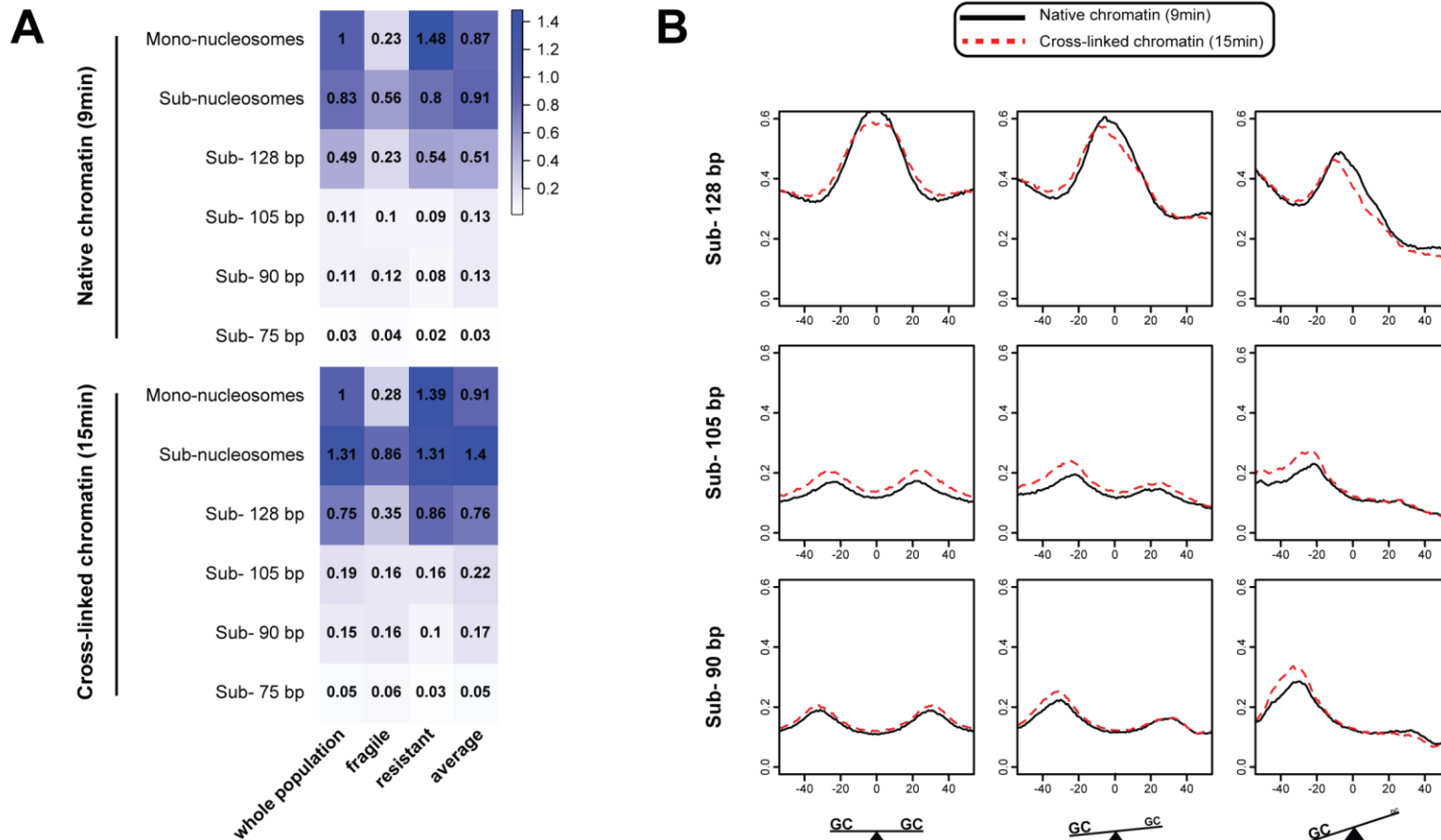
Moreover, in Ngo and colleagues measurements, the TA 10 bp periodicity played a crucial role in determining from which side nucleosomes unwrapped. Similarly, this study investigated the role of the DNA sequence in the genome-wide asymmetric MNase digestion within the nucleosome by taking in consideration local variations in GC content between the two nucleosomal halves. All called nucleosomes were divided by their GC content ratio between the two halves: 1) equal ratio, 2) slightly asymmetric, and 3) highly asymmetric (Figure 7.C). For each group, a fragment-centered V-plot was created to check where the sub-nucleosome fragment centers fell in relationship to the nucleosome dyad. When the GC content was equal between the two halves, a stochastically determined asymmetric MNase digestion occurred. However, when the GC content ratio was skewed, MNase digestion proceeded preferentially from the half with lower GC content, and this preference was stronger in higher GC ratios between the two halves. These observations were accentuated in prolonged MNase digestion levels and with cross-linked chromatin (Figure S3). Therefore, our results are in accordance to

Ngo and colleagues, and indicate that asymmetry in the underlying GC content of the nucleosomal DNA lead to an asymmetric MNase digestion within the nucleosome, likely driven by different rates of spontaneous unwrapping between the two halves.

Ngo and colleagues also demonstrated that asymmetric unwrapping further stabilizes DNA–histone interactions of the nucleosome side that remains wrapped. Therefore, it is possible that the tighter half of the nucleosome is more prone to be cross-linked, decreasing its probability to unwrap. To test this hypothesis, native and cross-linked chromatin preparations from prolonged, but similar, digestion levels were compared. Indeed, the fold changes of sub-fragment coverage against the mono-nucleosomal one were 0.83 and 1.31 (whole population column) in native and cross-linked chromatin, respectively. Thus, the protection of sub-nucleosomes is higher in cross-linked than native chromatin (Figure 9.A). Remarkably, the increase in protection induced by cross-linking was not equally distributed along the nucleosomal DNA. Mapped sub-fragments centers around nucleosomes divided as above (GC content ratios between halves) showed clear differences between native and cross-linked preparations (Figure 9.B). Although a great overlapping between the two chromatin preparations was present for the 128 bp peak, the 105 and 90 bp peaks showed higher protection in cross-linked chromatin. This further protection showed symmetry when nucleosomes with similar GC content between the halves were considered, whereas it was asymmetrical in nucleosomes with asymmetrical GC content.

Furthermore, considering the asymmetric MNase digestion within the nucleosome, sub-nucleosomal fragments ends can be located at the entry/exit of the nucleosome, and consequently the location of the cut within the nucleosome can be inferred (Figure 7.B). As a result, intra-nucleosome cuts for fragments of 128, 105 and 90 bp mapped at  $\pm$  55, 32 and 17 bp from the nucleosome dyad, respectively. Those nucleosomal coordinates are surrounded by regions of higher strength in DNA–histone interactions, as measured by nucleosome mechanical unzipping (Hall et al., 2009). This observation suggests an overlapping between MNase cuts within the nucleosome and the energetic diagram of the DNA-histone interactions.

# Figure 9



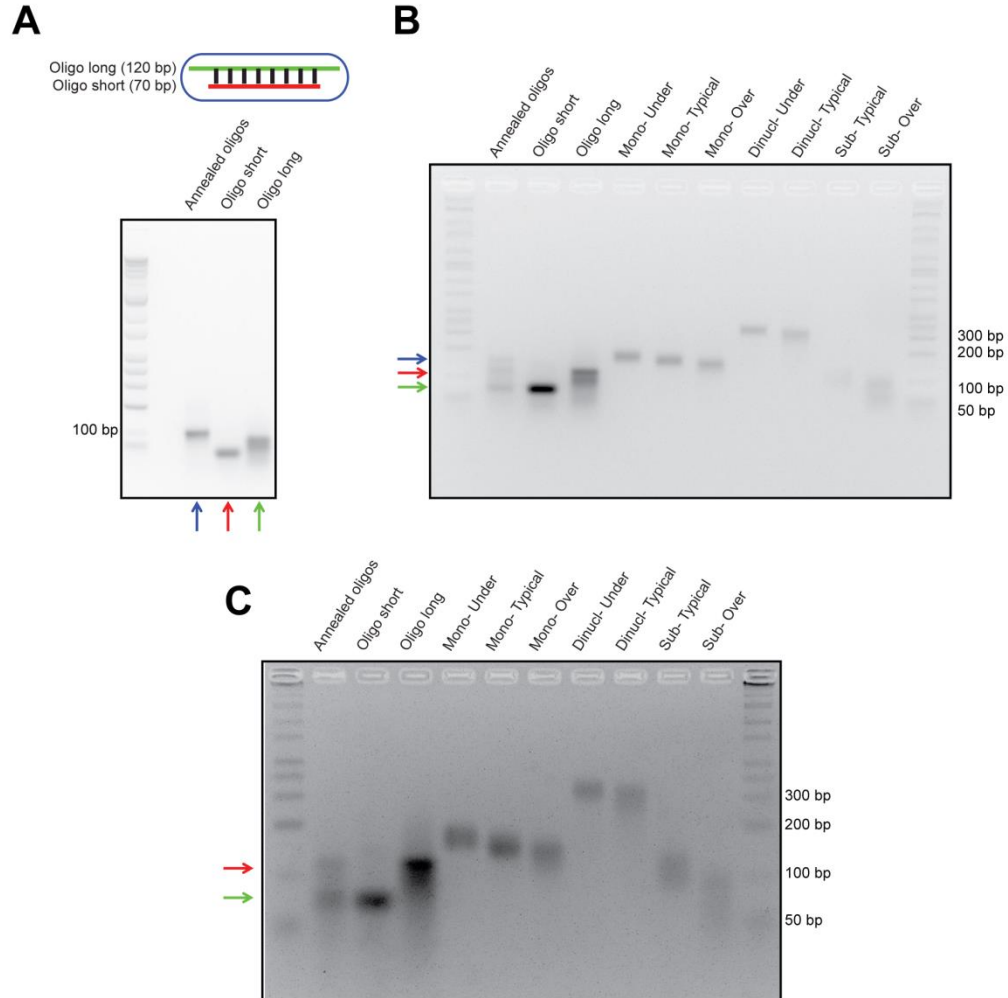
**Figure 9: Cross-linking effect on MNase digestion within the nucleosome.** (A) Coverage calculated from dyad occurrences of all called nucleosome through nucleR (threshold = 0.55; whole population) or of all nucleosome populations for mono- and sub-nucleosome fractions. In each chromatin preparation (native 9 min and cross-linked 15 min) data were normalized (divided by) against the coverage of the mono-nucleosome fraction over the whole population. (B) Horizontal slices of the V-plots derived from Figure 7.C for native chromatin (unbroken black lines) and from the second line of Figure S3 for cross-linked chromatin (dashed red lines). The slices are delimited by the sub-nucleosome peaks from the *in silico* size-selection: sub- 128 bp = 118-139 bp; sub- 105 bp = 100-110 bp; sub- 90 bp = 80-97 bp

To strength our hypothesis that asymmetric MNase digestion within the nucleosome is mostly driven by spontaneous unwrapping, frequency of single strand cuts within the nucleosomal DNA was measured. Indeed, it is known that MNase breaks on the DNA are the result of single nicks on both strands that are closer to each other. Therefore, MNase digestion on nucleosomal wrapped DNA should produce single nicks on the same strand with higher frequency than on unwrapped DNA, since the latter is protein-free and consequently more subjected to double strand breaks. All analyzed samples (dinucl-, mono- and sub-nucleosome fractions of all digestion levels) were run in a denaturing alkaline agarose gel (Figure 10). As a positive control, DNA oligomers of different length (120 and 70 bp) were designed, so that their annealed product was a dsDNA with a distinguishable size compared to the two ssDNAs (Figure 10.A). By using only the loading buffer as denaturizing agent, which does not allow a complete denaturation of dsDNAs, three bands in the positive control were observed, corresponding to the annealed product and the two ssDNAs (Figure 10.B). No variations in size distribution were observed for the chromatin samples in this condition. Instead, by using both loading and running buffer as denaturizing agents, permitting the complete denaturation of dsDNAs, the positive control was completely denaturated with only the two ssDNA bands visible (Figure 10.C). In this setup, none of our analyzed samples showed a different size distribution. These experiments suggest that nucleosomal fragments containing single nicks on one strand are rare in chromatin digestion. Conversely, single nicks on both strands are more common, indicating digestion of unprotected DNA as a more plausible explanation than digestion on bound DNA to generate sub-nucleosomes.

Taken together, these results indicate that the sub-nucleosome distribution only partially overlaps with the 10 bp periodicity of the nucleosomal DNA. More specifically, a pronounced discrepancy is present in the step from 105 to 90 bp, indicating that MNase does not only cut in each helical turn, as it would if MNase digestion proceeds on bound DNA. This is further confirmed by the overlapping between the sub-nucleosome distribution and binding energy map of DNA-histone interactions, suggesting that MNase can capture some of the binding energy preferences within the nucleosome. These evidences indicate the possibility for MNase to be used as a tool to evaluate binding energy between DNA and histones. Moreover, our analysis on asymmetric MNase digestion within the nucleosome suggests that the GC content ratio between

the two nucleosome halves could determine differences in binding energy and generate differential rates of asymmetrical unwrapping between the two halves *in vivo*.

## Figure 10



**Figure 10: MNase single nicks analysis.** (A) A positive control was designed through annealing of two ssDNA oligos of different length. The annealed dsDNA and the two oligos show different length in a TAE 3% agarose gel. (B) All differential MNase-seq samples were run in a TAE 3% agarose gel with an alkaline loading buffer, which permits a partial denaturation of dsDNAs, as shown by the positive control. (C) All differential MNase-seq samples were run in an alkaline 3% agarose gel with an alkaline loading and running buffer, which permits a complete denaturation of dsDNAs, as indicated by the positive control. Blue, red and green lines indicate presence and location of annealed and single strand oligos from the positive control.

#### **4.2.1 The +1 nucleosome of BP genes is asymmetrically digested by MNase**

To ascertain the biological relevance of the asymmetric GC content signature, in the next step, it was searched among nucleosome positions. Strikingly, the +1 position in BP genes was characterized by an asymmetrical GC content, with the proximal half to the TSS having a lower GC content than the distal half. Prompted by this observation, MNase digestion within the +1 position was characterized by mapping MNase cuts frequency. To also achieve an accurate investigation independent on activity components, each promoter cluster (NP, BP directional and BP divergent) was divided in active and inactive gene based on the expression level.

In BP divergent expressed genes, a higher frequency of sub-nucleosome cuts was observed within the proximal half to the TSS than the distal half (Figure 11.A). Moreover, sub-nucleosome cuts tended to proceed deeper within the nucleosome from the nucleosome entry, and to be more aligned with mono-nucleosome cuts at the nucleosome exit. These observations can be found also in the not expressed group. In this last case, the asymmetry of sub- typical- MNase cuts was even more accentuated and less likely to be influenced by non-nucleosomal protection, such as paused Pol II and other DNA binding proteins (Figure 11.B). Curiously, sub- typical- MNase cuts within the +1 nucleosome resembled the sub-nucleosome fragment distribution previously described, namely a continuum distribution with spikes around the length of 128, 105 and 90 bp. To take into account MNase sequence bias, cuts frequency from gDNA short was also plotted. Its distribution was not asymmetrical and the spikes were shifted or absent compared to sub- typical-. Therefore, the asymmetry in MNase cuts observed within the +1 position was not driven by the sequence bias, but from chromatin properties.

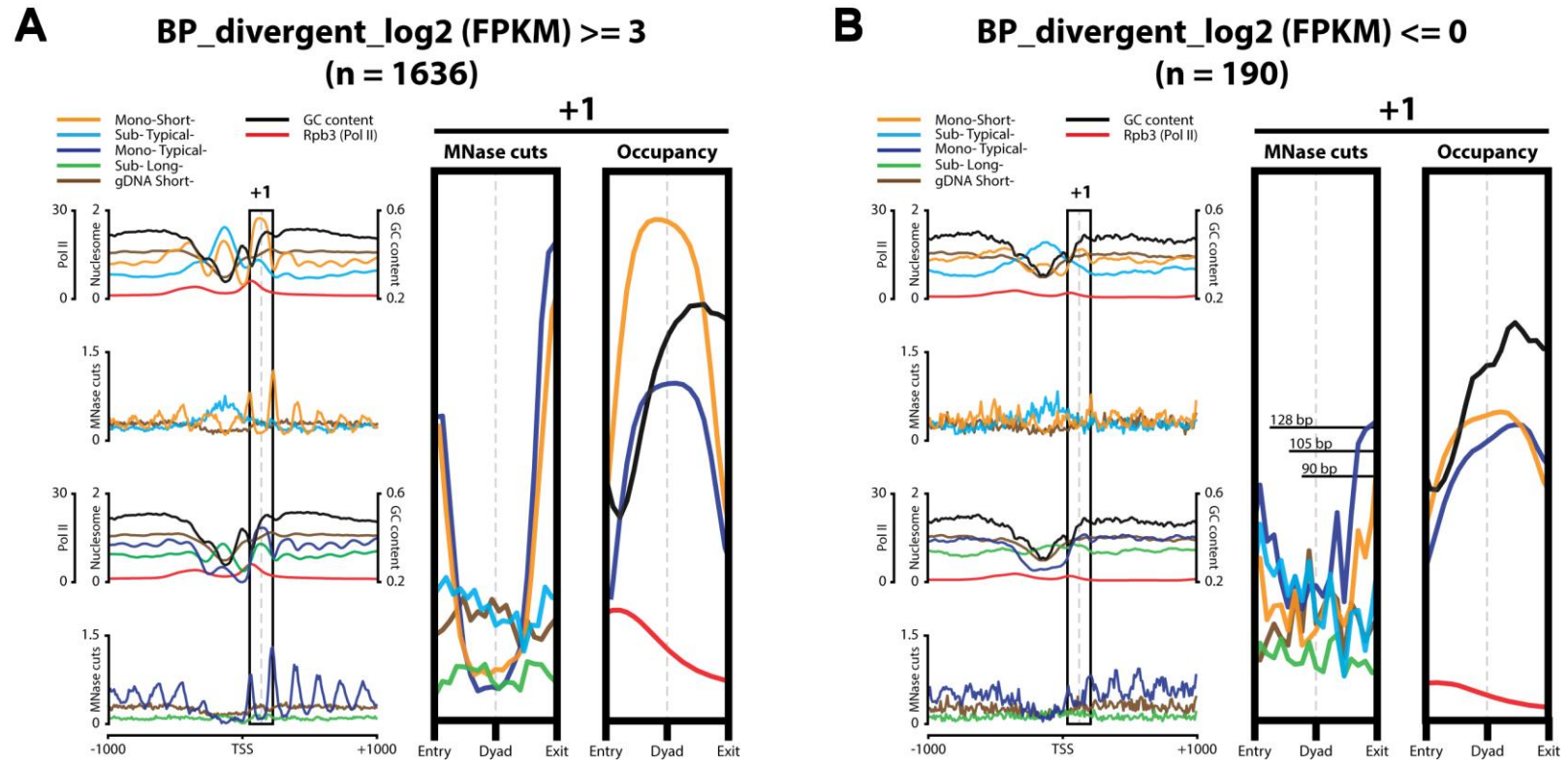
Asymmetric digestion was also visible in BP directional promoters, although not expressed genes showed a less pronounced pattern of spikes. This is most likely due to increased nucleosome fuzziness (Figure S4). By contrast, the +1 nucleosome of NP genes showed an underlying symmetrical GC content. Indeed, in the inactive NP genes group, sub-nucleosome cuts were symmetric within this nucleosome position. Instead, they were asymmetric when active genes were considered (Figure S5). The latter case could be explained by the strong occupancy of the paused Pol II that collides with the proximal half of the +1 nucleosome. Pol II offers a reduced protection against MNase than nucleosome and could induce higher local chromatin accessibility within its occupied region.



## *Results*

In conclusion, an asymmetric MNase digestion within the +1 nucleosome was observed in BP genes. It could be intriguing to hypothesize a sequence landscape of BP promoters evolved to facilitate the Pol II overcoming of the +1 nucleosome barrier by directing a spontaneous unwrapping from the proximal half.

## Figure 11



**Figure 11: MNase digestion within the +1 nucleosome (BP divergent genes).** (A) On the right composite plots around BP divergent (distance between TSSs of 200-500 bp) expressed ( $\log_2(\text{FPKM}) \geq 3$ ) genes. In brackets n is the gene count of the cluster. The first and third lines report mean coverage for Pol II and for some differential MNase-seq tracks. The same lines also report the GC content (25bp bins) profile. The second and fourth lines report the normalized MNase cut frequencies of the differential MNase-seq tracks. The +1 nucleosome position is graphically depicted to include the area surrounded by the first two mono- short- MNase cut peaks downstream of the TSS. On the left all MNase cut frequencies and the mean coverage of some tracks within the +1 nucleosome are enlarged for a better visualization. (B) Same plot of (A) around BP divergent not expressed ( $\log_2(\text{FPKM}) \leq 0$ ) genes, in order to remove the influence of activity components in the analysis of the MNase cut frequencies within the +1 nucleosome. Note the asymmetric MNase cut frequency of sub- typical- within the +1 nucleosome. Identical results were obtained with expressed gene shuffled and reduced in count to match the no expressed gene cluster size.

### 4.3 **PART II – Differential MNase-seq**

The digestion level bias mostly affects nucleosome occupancy and is exacerbated by the commonly used mono-nucleosome size-selection. If the mono-nucleosome fraction is imagined as a storage warehouse, its nucleosomal content can be considered as the inventory. The inventory continuously changes due to inflows (that is, the nucleosomal content from upper fractions) and outflows (that is, the nucleosomal content that goes to lower fractions or is digested away). MNase-seq with one digestion level can be considered as a picture of the inventory in a precise moment. Unfortunately, MNase digestion is not a homogenous process because inflows and outflows continuously change themselves and the inventory. Consequently, differential MNase-seq, namely a MNase-seq performed with multiple digestion conditions, is a better tool for analyzing the entire warehouse dynamics by taking into account also the time frame.

For instance, the digestion level bias can be monitored by looking at the canonical nucleosome pattern (Figure 2.B). Around BP promoters, -1, -2 and -3 nucleosomes showed a linear reduction in occupancy with the increase of the digestion level. The +1 nucleosome also showed a similar, but reduced, trend. Instead, a more complex behavior characterized the +2 nucleosome. Its occupancy firstly increased from mono- short- to mono- long- (15'), then decreased in prolonged digestion levels. Finally, further downstream nucleosomes (+3, +4 and so on) were resistant to MNase digestion, namely their occupancy linearly increased with the increase of the digestion levels.

Thus, differential MNase-seq is not a homogenous process along the chromatin and each nucleosome is characterized by its own sensitivity to MNase digestion. In this part, we used our differential MNase-seq setup to develop a new method that can cluster nucleosomes based on their MNase sensitivity and to explore the features behind the distinction between fragile and resistant nucleosomes.

### 4.3.1 Dinucl- and sub-nucleosomes provide new information on nucleosome landscape

To study the complexity in nucleosome occupancy with regard to digestion levels, inflows and outflows were investigated, namely the nucleosomal content from di-nucleosomes and sub-nucleosomes around CREs.

BP directional genes showed high occupancy of sub-nucleosomes upstream of the TSS, precisely at the -1, -2 and -3 positions. More importantly, the occupancy ratio between sub-typical- and sub-long- favored the former, indicating higher nucleosome accessibility to the MNase (Figure 12.A). At the position 0 a peak for sub-typical- was observed. However, neither the other samples nor our histone tail PTMs pull-downs can confirm this observation (Figure 22). For such a reason, this signal cannot be fully assigned as nucleosomal, but at the same time the presence of non-canonical nucleosome particles cannot be totally excluded. At the +1 position, sub-nucleosomes showed relative high occupancy. In contrast to the upstream region, downstream of the TSS the ratio between sub-typical and sub-long- favored the latter, suggesting a reduction in nucleosome accessibility which leads to an increased resistance against MNase digestion within the nucleosome.

Di-nucleosomes showed higher occupancy downstream of the TSS. The occupancy ratio between dinucl-short- and dinucl-typical- changed from the +1 and +2 nucleosomes, in which short-digestion showed higher signal, to the downstream nucleosomes, in which typical-digestion signal was predominant. This indicates that +1/+2 di-nucleosomes are located in regions with higher local chromatin accessibility compared to nucleosomes located within the gene body. The NDR proximity surely can play a role in that. Similar features can be retrieved in low expressed or shut down genes. In this case spatial resolution was more difficult to achieve due to the lack of any nucleosome pattern.

Similar observations can be made for the BP divergent cluster (Figure 12.B). However, higher chromatin and nucleosome accessibility at the -1 nucleosome position was detected in both active and inactive genes. Therefore, the intergenic region between divergent genes is characterized by a remarkably strong MNase sensitivity.

NP genes were characterized by broader nucleosome accessibility upstream of the TSS, covering almost 1Kb (Figure 12.C). In the -130/+70 bp window around the TSS, nucleosome accessibility showed strong dependency on the gene activity. Therefore, NP promoters seem to have an activity-dependent nucleosome organization that affects mostly the TSS and a large

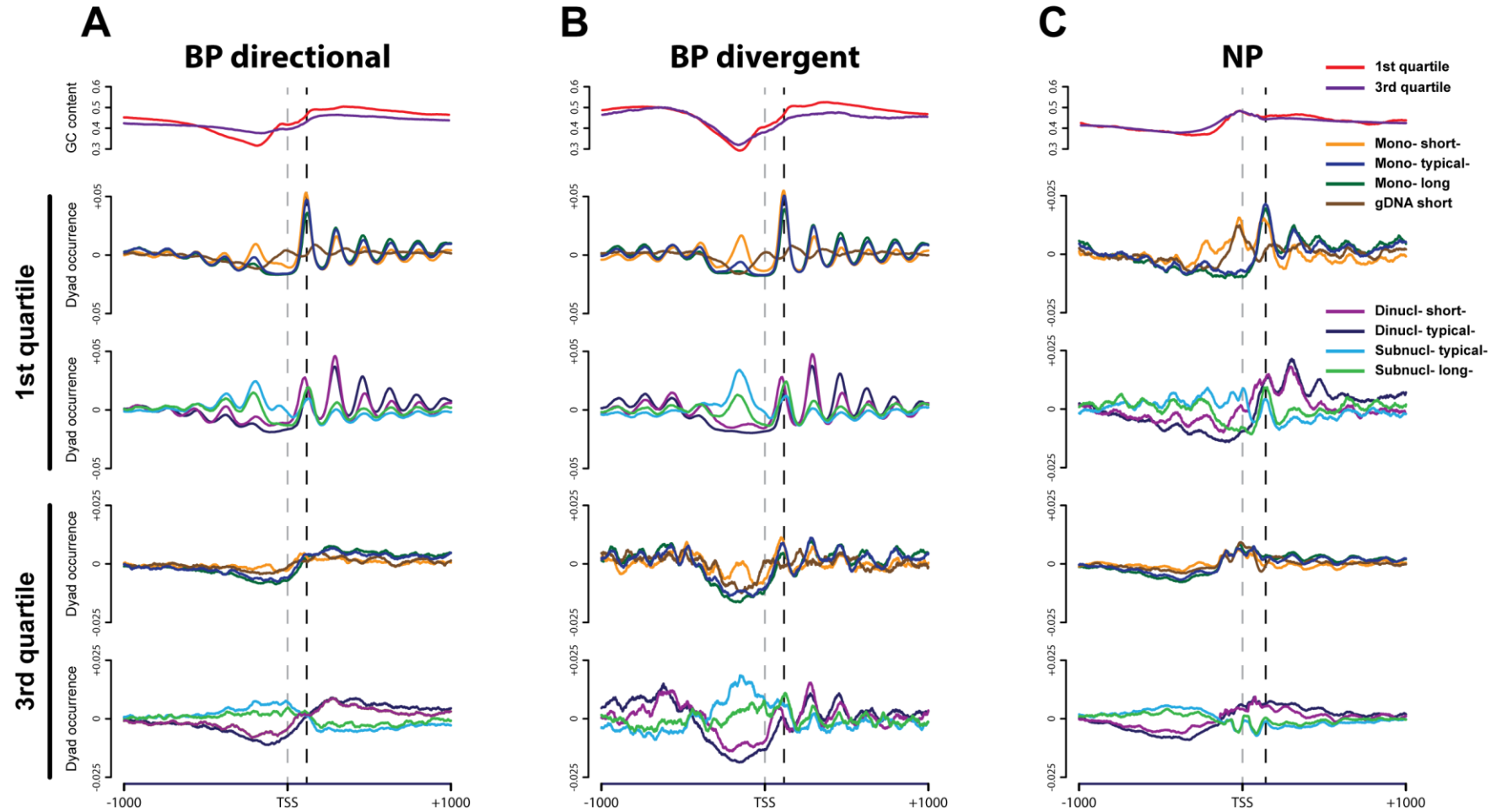
region upstream of it. Moreover, the downstream region is less structured and lacks a proper nucleosome pattern even in active genes.

In the surrounding 200 bp around DHS peaks a strong MNase sensitivity was detected in active enhancers (OSC/S2 shared and 1<sup>st</sup> quartile) (Figure 13.A). More specifically, higher occupancy of mono- short- over other mono-nucleosome samples was recorded, as well as higher sub- typical- over sub- long- and dinucl- short- over dinucl- typical-. DHSs peaks were also surrounded by two well positioned nucleosomes with a reduced but still strong MNase sensitivity. By contrast, the opposite situation was observed in further nucleosomes. Interestingly, inactive enhancers in S2 cells (OSC unique) showed higher and lower occupancy for di-nucleosomes and sub-nucleosomes, respectively. Nevertheless, the -2/-3 and +2/+3 nucleosomes depict the opposite situation. These findings suggest that inactive enhancers still preserve high local chromatin accessibility at their peak, but only surrounding nucleosomes are characterized by stronger inclination to be digested inside by MNase.

On the TTSs, a strong NDR was detected in all differential MNase-seq tracks, except from sub-nucleosomes, which show two peaks directly surrounding the TTS instead (Figure 13.B). Further investigations are necessary to clarify the nature of this signal. It could derive from nucleosomes strongly sensitive to MNase digestion, from non-canonical nucleosome particles or it could be an artifact of MNase digestion.

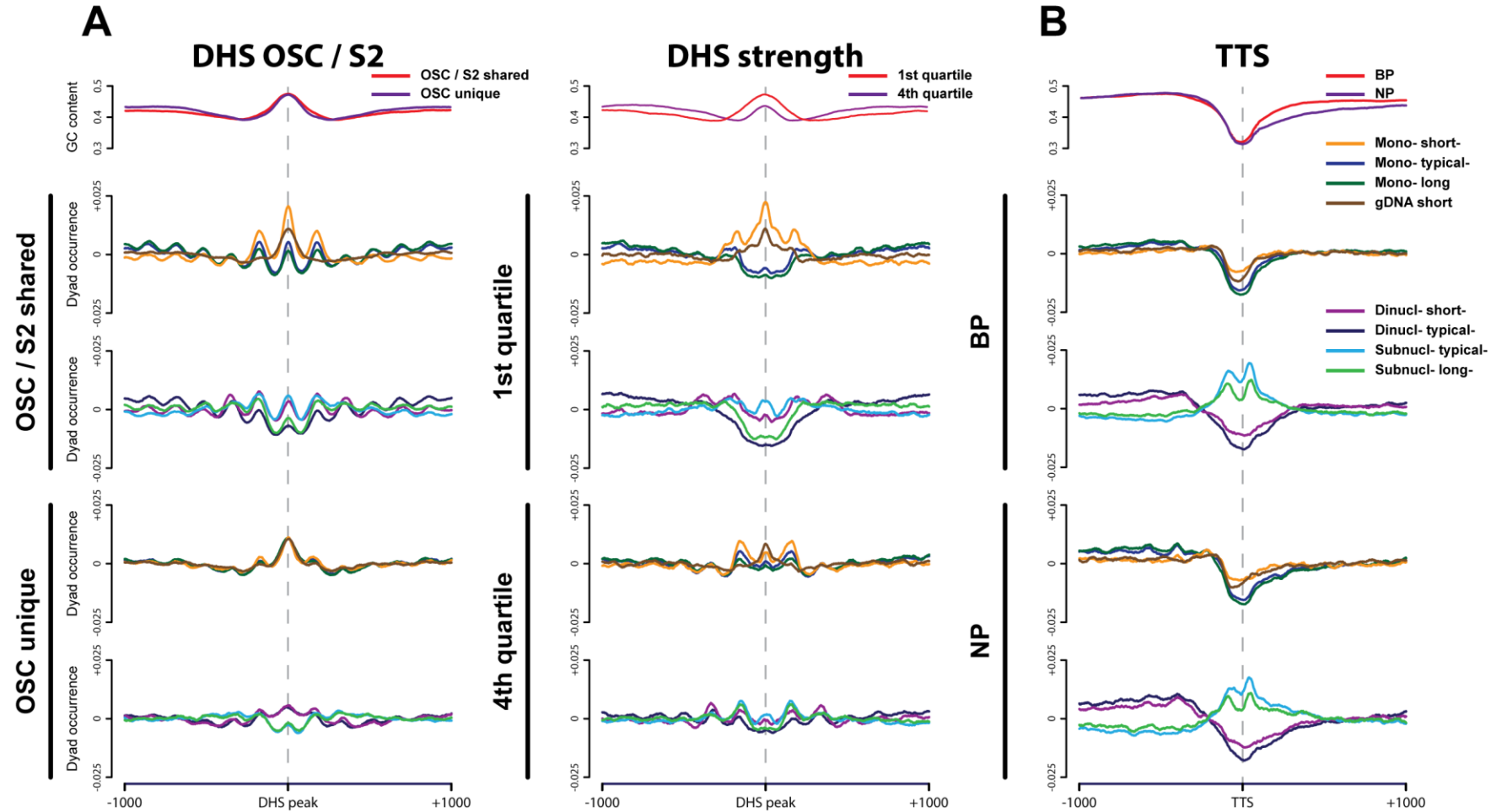
In conclusion, our data reveal that analyzing inflows and outflows of the warehouse can be useful to retrieve new information regarding chromatin features around CREs. Specifically, this approach can be relevant to distinguish nucleosomes not only based on their MNase sensitivity, but also to measure local chromatin and nucleosome accessibility to MNase. In the non-homogenous MNase digestion, nucleosomes located in more accessible regions of the chromatin tend to move earlier from the chromatin or oligo-nucleosomes to the mono-nucleosome fractions. In addition, nucleosomes characterized by weak DNA-histone interactions, which could lead to higher nucleosome accessibility within the nucleosome, tend to move earlier from the mono- to the sub-nucleosome fractions.

# Figure 12



**Figure 12: Differential MNase-seq around promoters.** In the first line it is shown the GC content profile (25 bp bin) for BP directional (A), BP divergent (B) and NP (C) promoter clusters. Each panel shows the first and third quartiles. Second and fourth lines are the smoothed and normalized nucleosome dyad frequency calculated from fragment occurrences for all mono- from differential MNase-seq and gDNA short samples. Third and fifth are the same for dinucl- and sub-nucleosome samples. Gray dashed lines are aligned to the TSS and black dashed lines are aligned to the +1 nucleosome dyad of the typical-digestion level.

## Figure 13



**Figure 13: Differential MNase-seq around DHS peaks and TTSs.** Subfigures matching those in Figure 12. (A) In the left panel, tracks around DHSs peaks uniquely active in OSC cells and shut down in S2 cells (second and third line) and around DHSs peaks shared between OSC and S2 cells (fourth and fifth line) are shown. In the right panel, tracks around DHS peaks divided in quartile of fold change against the input are shown (first and fourth quartiles). (B) Tracks around TTSs of BP (second and third lines) and NP (fourth and fifth lines) genes are shown.

### 4.3.2 Differential MNase-seq can distinguish nucleosome populations

As mentioned in the previous paragraph, new information can be gained through dinucl- and sub-nucleosome mapping. In a successive step, we asked how the fractions relate to each other throughout a differential MNase-seq. More specifically, correlation among all samples was performed through a PCC method.

Interestingly, longer fractions revealed higher correlations with shorter fractions of the successive digestion level (red arrows in Figure 3.D) than shorter fractions of the same digestion level (black arrows in Figure 3.D). For instance, sub- typical- had a correlation of 0.21 with mono- short- and a lower correlation of -0.02 with mono- of the same digestion level. This observation indicates the presence of distinct nucleosome populations that move from upper to lower fractions throughout a differential MNase-seq. Several features could contribute to the nucleosome population distinction, such as underlying DNA sequence, histone variants, histone PTMs, chromatin accessibility and active ATP-consuming mechanisms.

The DNA influence was investigated as di-nucleotide frequencies into the sequenced fragments (Figure 14.C). Each fraction (dinucl-, mono- and sub-) showed an increase in SS and a decrease in WW di-nucleotides, as well as a more pronounced 10 bp periodicity, with the increase of the digestion level. The same scenario occurs when fractions of the same digestion level were compared. High GC content and a more pronounced 10 bp periodicity are correlated with higher probability for a DNA sequence to be packed in a nucleosome. GC content of the underlying DNA sequence seems to capture the population shifting from longer to shorter fractions. Indeed, PCC analysis with GC content bins revealed an increment in GC content in each fraction with the increase of the digestion level (Figure S2.C). For instance, sub-nucleosomes showed a correlation of -0.37, -0.17, 0.23 and 0.51 for typical-, long- (15'), long- (9') and long- (30'), respectively. Therefore, increase of the digestion level leads to nucleosomes with higher GC content to be digested in sub-nucleosomes.

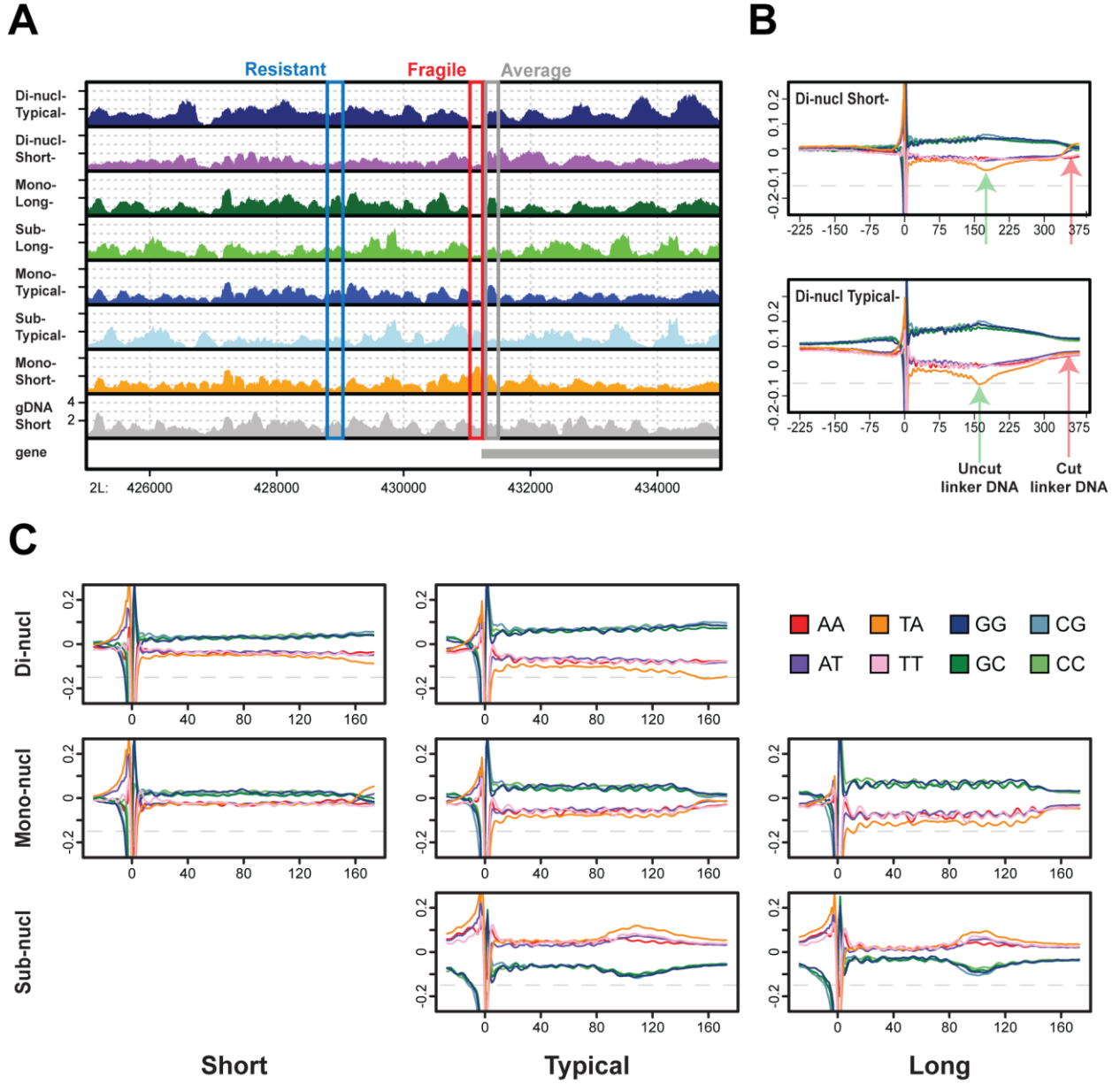
Given the unidirectional increase in GC content and 10 bp periodicity throughout a differential MNase-seq, we then asked whether these DNA features can be used to evaluate the digestion level of a chromatin digested sample. In this regard, di-nucleotide frequencies of chromatin samples prepared differently were compared: MNase digested native (9' and 30') against MNase digested cross-linked chromatin (15' and 45'). Similar digestion levels were obtained experimentally through the evaluation of the DNA signal from the agarose gel and by



calculating signal ratios among fractions. Both comparing native 9' to cross-linked 15' (Figure 6.C) and native 30' to cross-linked 45' (data not shown) pairs resulted in similar di-nucleotide frequencies in their mono- and sub-nucleosome fractions. To confirm that similarity in DNA features also means similar nucleosome population content, paired digestion levels were analyzed by mapping and genome-wide correlation. Interestingly, digestion level pairs greatly overlapped around BP promoters showing identical nucleosome positioning and occupancy (Figure 6.D). The overlapping in occupancy was particularly important to achieve given its strong dependency on the digestion level bias. Furthermore, digestion level pairs also showed high PCC values between fractions (around 0.88 for mono- pairs) (Figure S2.C).

Collectively, these results indicate that nucleosome populations can be distinguished throughout a differential MNase-seq. They tend to move from longer fraction to shorter ones with the increase of the digestion level. Thus, the key in the population distinction resides in the transitions among fractions. Moreover, DNA sequence plays a role in the nucleosome population distinction and it can be also considered as a metric to evaluate both nucleosome population content and digestion level bias within a MNase digested chromatin sample.

# Figure 14



**Figure 14: Genomic screenshot and DNA features of differential MNase-seq samples.** (A) Genomic tracks for all differential MNase-seq samples and gDNA short in an example 10 Kb region of chromosome 2L. Common profiles for fragile, average and resistant nucleosome populations are marked by colored boxes. (B) Smoothed di-nucleotide enrichments over the di-nucleosome fragments of short- and typical- digestion levels, showing divergent features between the undigested and digested linker DNAs. (C) Smoothed di-nucleotide enrichments over the nucleosome region for all differential MNase-seq samples (see Figure 4.C). The panel arrangements matched the one showed in Figure 3.A. The gray dashed lines are introduced to aid a better visualization over the enriched and depleted sequence features, which depend on the digestion level and extracted fragment length.

### 4.3.3 Nucleosome populations can be identified within one digestion level

For each nucleosome position, occupancy can change throughout a differential MNase-seq. These changes are mostly linear against the digestion level, but they can also be complex, such as the +2 nucleosome downstream of the TSS (paragraph 4.3). Occupancy changes can be assigned to variation in nucleosome population content driven by the transition from longer to shorter fractions. To distinguish nucleosome populations, three simple definitions can be used:

- 1) Fragile nucleosomes: whose occupancy significantly decreases with the increase of the digestion level;
- 2) Average nucleosomes: whose occupancy does not change significantly throughout a differential MNase-seq;
- 3) Resistant nucleosomes: whose occupancy significantly increases with the increase of the digestion level.

These nucleosome behaviors are shown in the genome-wide screenshot of Figure 14.A. Fragile nucleosomes (highlighted in red) showed high occupancy in mono- short- and in sub-nucleosome of the successive digestion level. Its occupancy strongly decreased or disappeared in the other samples. Average nucleosomes (highlighted in grey) showed no significant fluctuations in occupancy throughout the samples. Resistant nucleosomes (highlighted in blue) showed low occupancy in mono- short- and sub- typical-, and increased occupancy in the other samples.

Next, a new strategy to call and distinguished nucleosome populations at genome-wide fashion was established.

For nucleosome calling, the package nucleR was used (Flores and Orozco, 2011) on mono-nucleosome samples from the three digestion levels. In the calling step, a pre-determined threshold value of 0.55 was applied to have roughly one nucleosome for each 200 bp. Results were consistent also with higher thresholds (0.9), which reduced the amount of fuzzy nucleosomes in the calling. Finally, called nucleosome dyads from the three digestion levels were merged into a single track (**whole population**).

For the nucleosome population distinction, the differential information among fractions, rather than digestion levels, was extrapolated. Specifically, the transition from mono- to sub-nucleosome was measured, since it can better capture nucleosome accessibility than local chromatin accessibility to MNase. Indeed, this transition is driven by the binding energy between histones and nucleosomal DNA. Fragile nucleosomes are the first to be digested inside

throughout a differential MNase-seq. Therefore, mono- to sub-nucleosome transition of the typical digestion level can better assess this behavior. In contrast, resistant nucleosomes are enriched in longer fractions. For a better comparison with the fragile definition, the same transition was considered, but in the opposite direction. Thus, the ratio in occupancy between mono- and sub- of the typical digestion was calculated for each called nucleosome. Nucleosomes were defined as **fragile** when  $\text{sub-nucleosome}^{\text{typical-}} : \text{mono-nucleosome}^{\text{typical-}} > 2$ ; as **resistant** when  $\text{mono-nucleosome}^{\text{typical-}} : \text{sub-nucleosome}^{\text{typical-}} > 2$ . Every other nucleosome not included in these definitions was defined as **average** (Figure 15.A). Using these definitions, nucleosomes were split into 7% fragile, 49% average and 44% resistant in our data.

In conclusion, we developed a new differential MNase-seq approach to call three nucleosome populations (fragile, average and resistant) that takes into account fragment size than digestion level. Specifically, the mono- to sub-nucleosome transition was used to distinguish nucleosome populations based on DNA-histone interactions strength. Remarkably, our method calls populations through only one digestion level, avoiding the tedious experimental praxis to carry out multiple digestion levels for each experiment or condition. Moreover, this method is flexible since ratios between mono- and sub-nucleosome occupancies can be modified to refine the nucleosome population distinction or multiple ratios can be used to increase the number of called populations.

#### 4.3.4 Nucleosome populations have distinct DNA features

Once nucleosome populations were clustered, their underlying DNA features were characterized by assessing di-nucleotide enrichments. Indeed, as mentioned before, DNA sequence can be considered as a metric to evaluate nucleosome population content and digestion level bias.

Compared to whole population, fragile nucleosomes were enriched for WW di-nucleotides and depleted for SS di-nucleotides. By contrast, resistant nucleosomes showed a reversed situation, whereas the di-nucleotide profile of average nucleosomes fell in between (Figure S6.B). To assess the variation in DNA sequence within each population, we sought to measure sequence features for each called nucleosome. A standardized di-nucleotide nucleosomal DNA PWM was computed using fragments from mono- typical- with a 14 bp trimming from either side to avoid MNase sequence bias at the cut sites. Subsequently, the obtained PWM was computed in relationship to a PWM density score for fragments from each

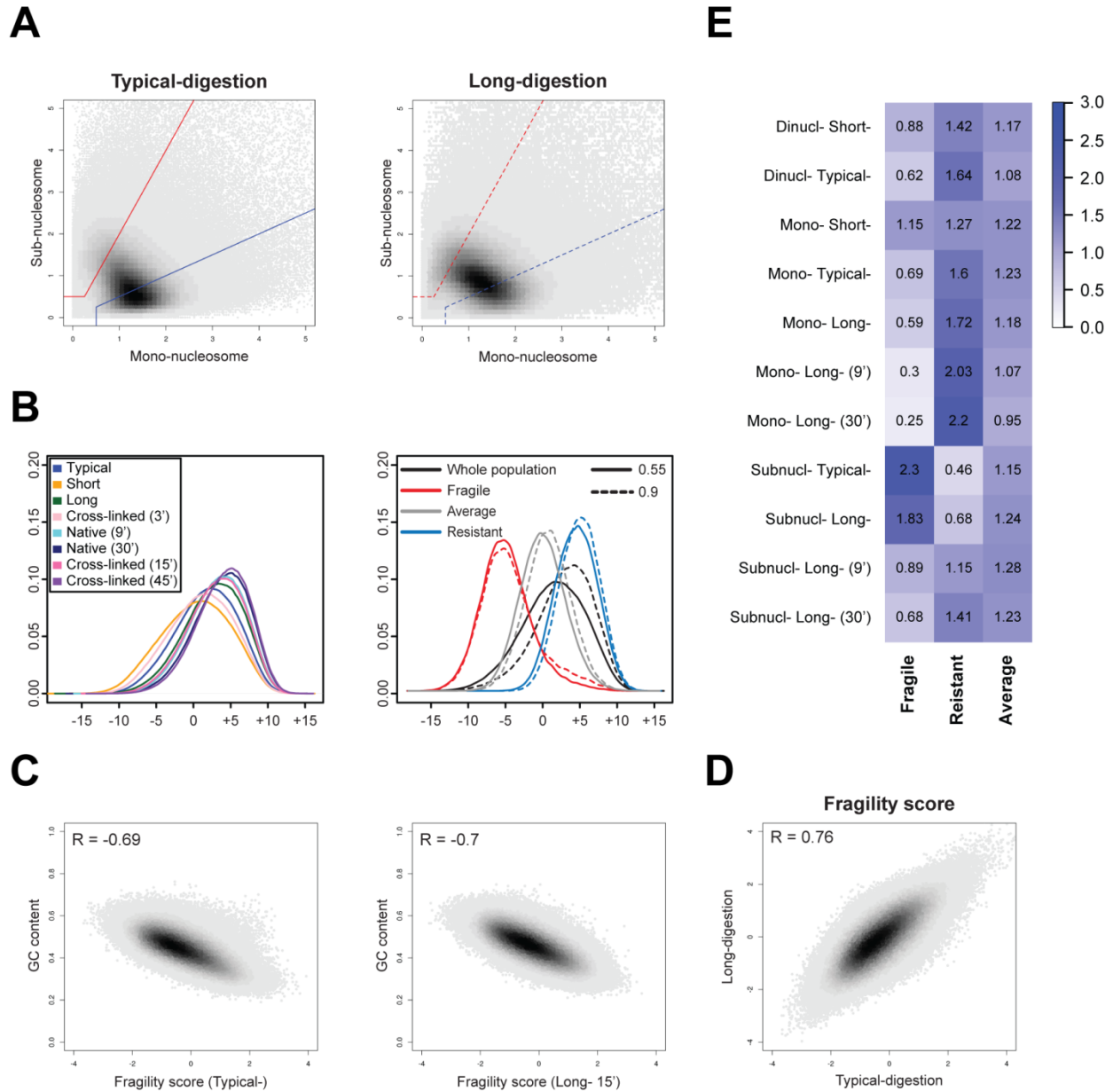
mono-nucleosomal sample and nucleosome population (Figure 15.B). In this way, fragments with a positive PWM score were characterized by a DNA sequence that favors nucleosome wrapping and vice versa for fragments with a negative PWM score. Interestingly, PWM score distributions of mono-nucleosome samples strongly overlapped, since they contained all three populations in different proportion. Moreover, these distributions moved toward positive PWM score with the increase of the digestion level.

In stark contrast, nucleosome population distributions were greatly separated. The fragile population was markedly shifted towards negative PWM scores, indicating enrichment for disfavoring DNA sequences to the nucleosome wrapping. The resistant population was shifted toward positive PWM scores, indicating enrichment for favoring DNA sequences to the nucleosome wrapping. Average population showed a distribution centered at 0. Furthermore, DNA sequence features underlying the three populations were identical when a reduce amount of fuzzy nucleosomes was included in the whole nucleosome population (nucleR threshold of 0.9). Nevertheless, the 0.9 whole population distribution was shifted toward positive values compared to the 0.55 one, indicating that a different selection of called nucleosomes was made. Thus, our approach is able to capture these populations independently on which nucleosomes are called.

Finally, the fragility score was significantly anti-correlated with the underlying GC content ( $R = -0.69$ ) (Figure 15.C), indicating its prominent role on the nucleosome population distinction. For such a reason, the GC content was highlighted in the successive nucleosome populations mapping.

In conclusion, nucleosome populations are characterized by distinct underlying DNA features. Moreover, the differences in DNA features observed in nucleosome populations are more accentuated than chromatin samples, suggesting that diverse DNA-histone binding energies correlate with the population behaviors throughout a differential MNase-seq.

# Figure 15



**Figure 15: Nucleosome populations are characterized by different sequence features.** (A) Scatterplot between mono- and sub-nucleosome coverage of called nucleosome for typical- (left panel) and long- (15') (right panel) digestion levels. The red and blue lines are the thresholds used to call fragile and resistant nucleosomes, respectively. (B) Sequence-feature score (1st-order Markov model) distribution of different nucleosome groups. The left panel shows the distributions for all mono-nucleosome samples, which are very similar. The right panel shows the distributions for nucleosome populations within the typical-digestion level. Nucleosome populations have distinct average scores and therefore sequence features. Unbroken and dashed lines in the same panel (continue in the next page) indicate the distribution of the nucleosome populations from nucleosomes called with two thresholds in

nucleR (0.55 and 0.9 respectively). (C) Anti-correlation between nucleosome fragility of typical-digestion (left panel) and long- (15') digestion (right panel) with the average GC content of the nucleosome region. (D) Scatterplot showing the significant correlation between the fragility score based on the typical- and long- (15') digestions. (E) Normalized average occupancy for differential MNase-seq samples over nucleosome populations.

#### 4.3.5 The one-digestion level approach is quite robust against digestion level

In a successive step, the robustness of our approach in nucleosome population calling was analyzed against the digestion level bias. For this purpose, the pipeline applied for the typical-population calling was used on the mono- to sub-nucleosome transitions from more prolonged digestion levels.

By comparing the two population sets, fragile nucleosomes from typical- and long- (15') digestion levels were highly correlated ( $R = 0.76$ ) (Figure 15.D). Moreover, the two fragile populations also shared pretty similar anti-correlation with the underlying GC content (Figure 15.C). Finally, similar di-nucleotide frequencies between the two sets of populations were also retrieved (Figure S6.B). Nonetheless, this robustness against the digestion level bias was reduced when more prolonged digestion levels were considered (9' and 30'). In these cases, fragile nucleosomes were characterized by a decrease in WW di-nucleotides, resistant nucleosomes by an increase in SS di-nucleotides and average nucleosomes by a more remarkable separation between WW and SS frequencies (Figure S6.B). As confirmation, PWM score distributions of populations derived from more prolonged digestion levels were also shifted toward positive values (Figure S6.A), indicating enrichment for DNA sequences that favor nucleosome wrapping. Finally, only long- (15') populations were highly correlated with the typical- ones genome-wide (Figure S6.C).

In conclusion, a one-digestion level approach for differential MNase-seq is quite robust against the digestion level bias, but only in a certain range that spans between the typical- and long- (15') digestion levels. This is most likely due to a higher probability for fragile nucleosomes to be digested away in more prolonged digestion levels. For such a reason, a protocol optimization to mimic our typical digestion level is recommended.

#### 4.3.6 Nucleosome populations are located differently around CREs

Subsequently, nucleosome populations were mapped around promoters, DHSs and TTSs, whose definition can be found in paragraph 4.1.2. Moreover, sequence and activity components were

also included in the nucleosome population mapping. Given the anti-correlation between fragility score and the underlying GC content, sequence components were considered as GC content profile. In parallel, activity components were considered by dividing genes in quartile of expression level and DHSs in quartile of signal strength.

Fragile and resistant populations mapped into separate locations in both BP directional and BP divergent promoter clusters. Fragile nucleosomes were enriched upstream of the TSS, whereas resistant and average nucleosomes were enriched mostly over the gene body (Figure 16.A & Figure 16.B). Specifically, in active genes, fragile nucleosomes were dispersed in a 500 bp window upstream of the TSS in BP directional, and they were focused at the -1 position in BP divergent. Interestingly, fragile and resistant nucleosome landscapes matched variations of the underlying GC content. Indeed, fragile nucleosomes of the BP directional cluster overlapped with the broad GC depletion upstream of the TSS, and the fragile -1 nucleosome of the BP divergent cluster exactly overlapped with the lowest point in the GC content. In parallel, variations of the resistant nucleosome frequency overlapped with similar variations of the underlying GC content downstream of the TSS. Separating promoters in quartile of expression level, directional promoters in the top 2 and divergent promoters in the top 3 quartiles showed a pronounced nucleosomal array, while the other quartiles were characterized by a diminished nucleosomal array. However, fragile and resistant nucleosomes retained their location, even if a decreased frequency of fragile nucleosomes was measured. Collectively, these data indicate that fragility and resistance are particularly carved in the DNA sequence of BP promoters. At the same time, activity components can affect fragility and resistance by amplifying their frequency in active and inactive genes, respectively. Interestingly, higher frequency of fragile nucleosomes was measured upstream of BP divergent promoters, even in shut down genes. Indeed, in this cluster, a stronger and sharper GC depletion was measured, indicating an even higher DNA-encoded -1 fragility.

To take in consideration some of the activity components, Henikoff's salt fractionated H3.3 and H2A.v pull-downs were measured around promoters (Henikoff et al., 2009) by taking into account two regions around the TSS for the -1 and +1 nucleosomes: (-150 +/-73) and (+120 +/-73), respectively (Figure S7.A). As expected, both histone variants were more deposited over high than low expressed genes. Interestingly, -1 and +1 nucleosomes were differently enriched for the two histone variants: H3.3 for the former and H2A.v for the latter. In addition, a stronger



enrichment for the low-salt fractionated chromatin (80mM) was measured for the -1 nucleosome. Indeed, low-salt fractions are enriched for low stable nucleosomes. Finally, differences in histone variant enrichment were measured between the two BP promoter clusters. Firstly, both histone variants were more deposited at the -1 and +1 positions for BP divergent and directional, respectively. Secondly, histone variant enrichment was found more dependent on activity in BP divergent. Collectively, these observations suggest for a more steady-state histone variant content in the intergenic region between divergent genes.

Next, similar analyses were performed on the NP promoter cluster. In this case, a completely different GC content profile, characterized by weaker fluctuations than BP clusters, was measured (Figure 16.C). Specifically, a shallower reduction in a 500-700 bp window upstream of and a peak directly on the TSS were observed. Therefore, in inactive genes, the TSS was covered by resistant nucleosomes, and in the upstream region a low frequency of fragile nucleosomes was found. In stark contrast, high frequency of fragile nucleosomes was measured in the same regions when active genes were considered. In conclusion, NP promoters show a DNA-encoded resistance on the TSS. Consequently, nucleosome fragility observed in expressed genes is mostly driven by active mechanisms.

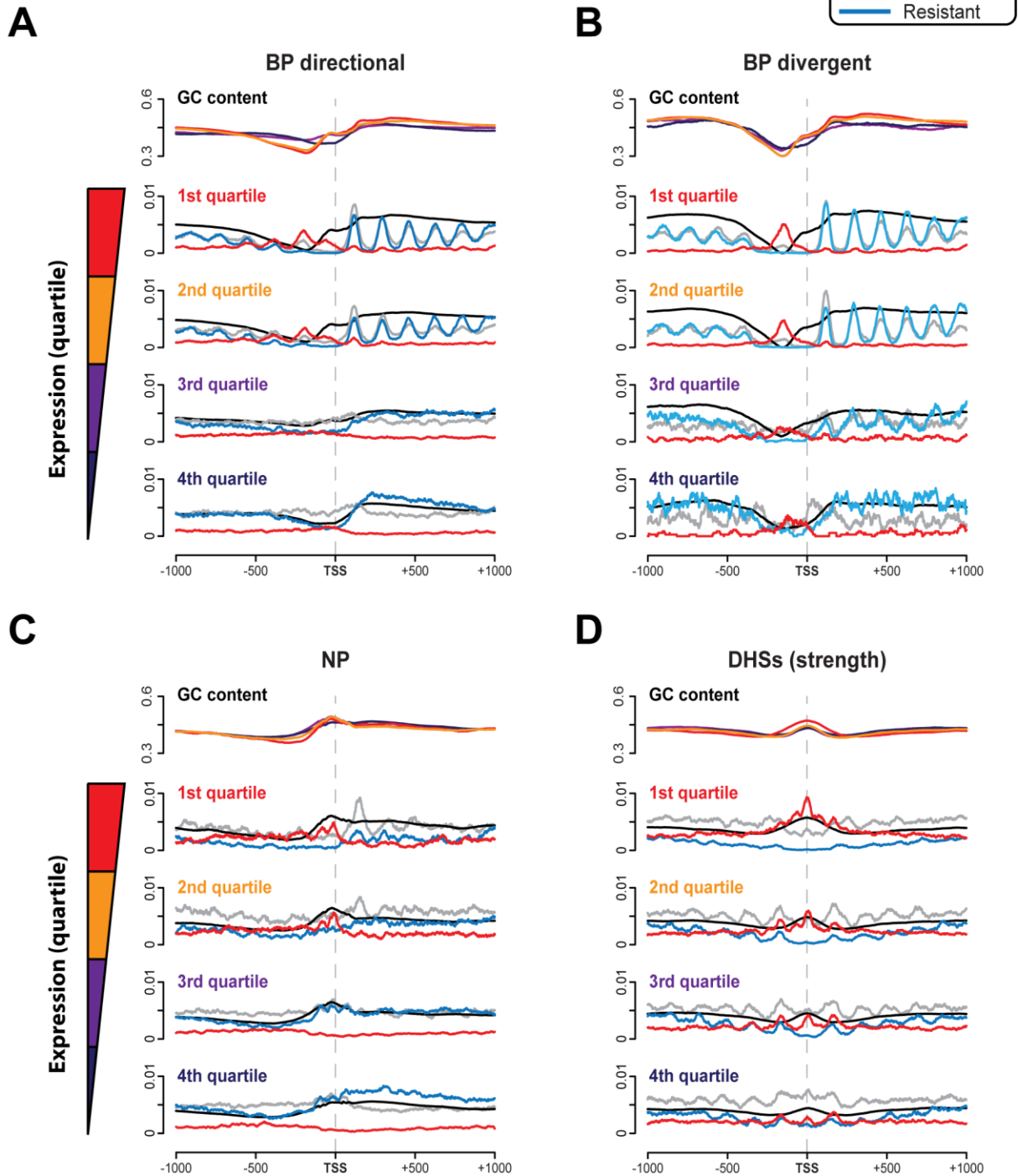
Strikingly, a GC content profile resembling NP promoters was found at the DHSs, namely a bump on the DHS peak symmetrically flanked by shallower reductions in GC content (Figure 16.D). Most importantly, the same sequence landscape was retrieved in DHSs totally inactive in S2 cells (OSC unique), indicating this GC content profile as a sequence signature of enhancers (Figure S7.C). As already observed in NP promoters, inactive enhancers in S2 cells were also covered by resistant nucleosomes at their peak and showed a slight increase in fragility in the surrounding. Consequently, nucleosome fragility observed at the DHS peaks in active enhancers (OSC/S2 shared and S2 unique) can be assigned as activity components driven. Furthermore, a quantitative correlation between fragility frequency and DHS openness was also observed (Figure 16.D). Indeed, the highest fragility at the DHS peak and broader distribution of fragile nucleosomes in the surrounding was measured in the first quartile. In parallel, fragility decreased with the decrease in DHS openness, until surviving only at the peak in the fourth quartile. Therefore, enhancers also show a DNA-encoded nucleosome resistance, and their activation leads to a drastic chromatin change and to a broad increment in fragility, which are correlated to the amplitude of the enhancer activation.

Next, the same histone variant pull-downs from fractionated chromatin were mapped around DHSs. H3.3, but not H2A.v, was found enriched in active enhancers. In addition, a moderate enrichment for H3.3 was also measured in the OSC unique cluster, indicating the priming of inactive enhancers (Calo and Wysocka, 2013) (Figure S7.B). Strikingly, active enhancers were more enriched in H3.3 from high-salt (150-600mM) than low-salt fractions (80mM). Therefore, in strong contrast to the -1 position upstream of the TSS in BP promoters, fragile nucleosomes within the DHSs were characterized by higher intrinsic stability. The DNA sequence could play a role in this difference, since nucleosomes within the DHSs are characterized by higher GC content, hence, higher intrinsic stability. This observation further confirms that fragility within the DHS is mostly activity, and not sequence, driven.

Finally, nucleosome populations were also mapped around TTSs, in which a strong depletion in GC content was measured. The GC valley progressively reached the genomic average within 1Kb into the intergenic region (Figure S7.D). Consequently, strong nucleosome depletion with a relative higher frequency of fragile nucleosomes was observed at the TTSs. Interestingly, a flatter GC content in the intergenic region was measured in NP genes. Concomitantly, a broader fragility was also found in the same region, indicating slight differences in chromatin structure around TTSs between NP and BP genes.

Taken together, nucleosome fragility and resistance are differently located around CREs. Moreover, fragility upstream of the TSS in BP promoters and resistance on the TSS of NP promoters and DHS peaks seem to be partially carved in the DNA sequence. In contrast, activity components drive the fragility on NP promoters and DHS peaks. Finally, the intergenic region between divergent genes shows a particular chromatin landscape, characterized by a stronger and sharper DNA-encoded fragility. Fragility and resistance dependency on sequence and activity components can also interplay. On one side, DNA-encoded fragility and resistance could aid or delimit active mechanisms. On the other side, active mechanisms could have been evolved to take advantage of the DNA-encoded fragility and resistance. In this regard, a detailed discussion follows in paragraph 5.2.3.

## Figure 16



**Figure 16: Nucleosome populations around CREs.** Smoothed profiles of nucleosome populations around BP directional (A), BP divergent (B) and NP (C) promoters. The top panel shows the GC content, followed by the profiles of fragile, average and resistant nucleosome populations. Promoters belonging to the four gene expression quartiles are shown as individual lines. (D) The same figure as other panels, but surrounding DHS peaks. The quartile separation is based on the fold-enrichment of the signal against the input within the DHS peaks.

#### 4.3.7 Alternative scores for nucleosome population distinction: mono- and oligo- based scores (monos and oligos)

This study focused on the transition from mono- to sub-nucleosome in order to develop a differential MNase-seq method able to discriminate fragile and resistant nucleosomes. In this way, nucleosome accessibility, and not local chromatin accessibility, to MNase was more taken into account. Subsequently, we asked what kind of information can be retrieved by considering other transitions. Specifically, the horizontal transition between mono-nucleosomes from different digestion levels and the vertical transition from oligo- to mono-nucleosomes were analyzed. These transitions mimic approaches used to study MNase chromatin accessibility in a plethora of species (Mieczkowski et al., 2016) and MNase sensitivity in *D.melanogaster* (Chereji et al., 2016), respectively.

Thus, two ratios were computed: monos and oligos. Monos was based on the mono-short- : mono- long- (15') ratio. In this case, nucleosomes were considered as fragile\* when  $\text{mono-nucleosome}^{\text{short-}} : \text{mono-nucleosome}^{\text{long- (15')}} > 2$ , and as resistant\* when  $\text{mono-nucleosome}^{\text{long- (15')}} : \text{mono-nucleosome}^{\text{short-}} > 2$ . Oligos was based on the dinucl- short- : mono-short- ratio. In this case, nucleosomes were considered as fragile\* when  $\text{mono-nucleosome}^{\text{short-}} : \text{di-nucleosome}^{\text{short-}} > 2$ , and as resistant\* when  $\text{di-nucleosome}^{\text{short-}} : \text{mono-nucleosome}^{\text{short-}} > 2$ . In both ratios, the average\* definition was kept as nucleosome being neither fragile nor resistant.

Applying the same pipeline used for the typical- populations, sequence features derived from the new fragile\* and resistant\* scores were analyzed by calculating the PWM score distributions. Monos and oligos distributions were found broader and characterized by a greater level of overlap (Figure S6.A), indicating a reductive ability to distinguish population based on the DNA sequence. Specifically, monos fragile\* and resistant\* distributions only partially resembled the typical- ones. Instead, a completely shifted distribution toward positive PWM scores was obtained for the monos average\* population. In parallel, oligos and typical- distributions were completely shifted against each other. Oligos fragile\* distribution was centered at 0 with half of the population characterized by positive PWM scores. Average\* distribution was also shifted toward positive PWM scores and the resistant\* distribution showed a long tail toward negative PWM scores. Furthermore, these discrepancies between typical- against monos and oligos populations were also mirrored genome-wide, in which low or no correlation was measured (Figure S6.C). Interestingly, these correlations were consistently lower

than populations calculated from mono- to sub-nucleosome transitions from different digestion levels.

In conclusion, mono- and oligo- based scores generate nucleosome populations that strongly differ from the typical- ones, both at the DNA sequence and genome-wide levels. Moreover, the distinction among monos and oligos populations is mostly driven by other properties than DNA sequence. Given the nature of monos and oligos transitions, local chromatin accessibility can be eligible as candidate.

#### **4.3.8 Mono- and oligo- scores better capture local chromatin accessibility**

To ascertain how monos and oligos populations differ from the typical- ones, they were mapped around CREs. Furthermore, to include the digestion level bias in such a comparison, populations obtained from the mono- to sub-nucleosome transition of a more prolonged digestion level (long- 15') was also considered (Figure 17).

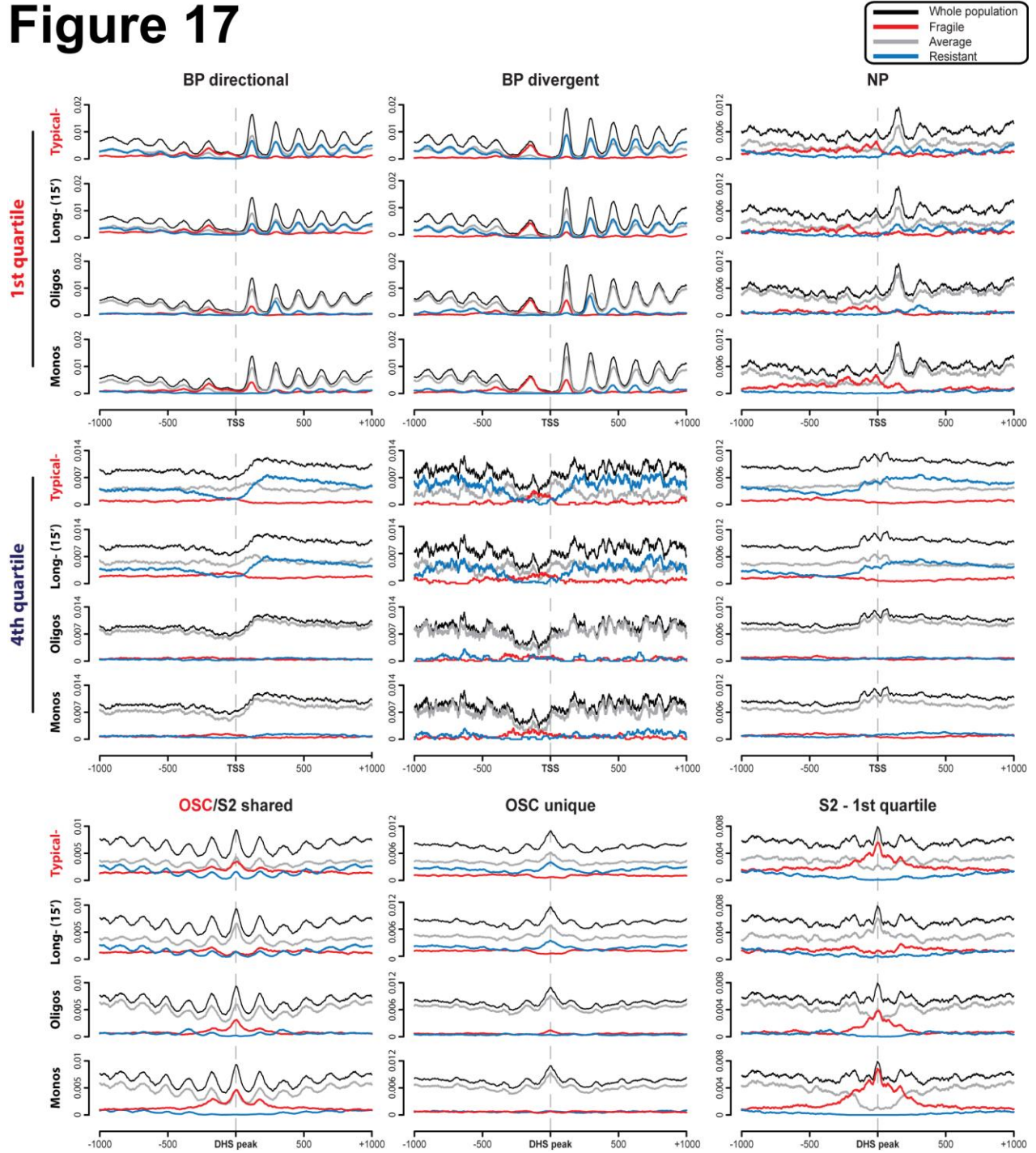
A great consistency between typical- and long- (15') populations was obtained, even if some differences were also detected. Specifically, a reduction in resistance at the +1 position in BP active genes was observed. At the same time, reduced fragility and resistance at the TSS of active and inactive NP promoters were detected, respectively. Finally, the most drastic change was measured on active enhancers, whose fragility was found strongly reduced. In all cases, fragile and resistant nucleosomes were replaced by average ones.

Collectively, these observations indicate that digestion level bias mostly affects the fragile population. This is particularly true for the activity dependent nucleosome fragility, as indicated within NP promoters and enhancers.

Regarding monos and oligo populations, more pronounced differences were observed against the typical- ones. For instance, resistant nucleosomes at the +1 position were replaced by fragile\* and average\* ones in active promoters. This replacement could derive from its higher local chromatin accessibility due to the NDR proximity. Furthermore, the resistance\* over the gene body was focused at the +2 nucleosome in oligos or it was strongly reduced in monos. Regarding the oligos score, it is possible that the resistance measures a second order of local accessibility, since dinucl- short- is still enriched for nucleosomes that derive from accessible regions of the chromatin. Regarding the monos score, the reduction in resistance can be explained by a reduced contribution played by the DNA sequence into the nucleosome

population distinction, since horizontal comparisons are not focused on the DNA–histone interactions strength. In inactive promoters, both oligos and monos fragile\* and resistant\* populations were completely undistinguishable. Indeed, these promoters share a reduced chromatin accessibility which cannot be captured by the two scores. Similar pictures and explanations were obtained when active and inactive enhancers were considered. Oligos resistance\* showed a second order of accessibility surrounding DHS peaks and monos fragility\* was increased in the same region due to its strong dependency on local chromatin openness. Once again, resistance on inactive enhancer was not captured from both scores.

In conclusion, other transitions than mono- to sub-nucleosomes seem to capture local chromatin accessibility. Indeed, they are unable to distinguish nucleosome populations in closed chromatin regions. In stark contrast, the DNA-histone binding strength distinguishes nucleosome fragility and resistance in any context, since it derives from single nucleosomes and not from the local chromatin accessibility.

**Figure 17**

**Figure 17: Nucleosome populations from alternative scores around CREs.** Subfigures matching those in Figure 16 to compare typical nucleosome populations (preferred in this study) against long- (15'), oligos and monos ones. Nucleosome populations are mapped around active (first line; first quartile) and inactive (second line; fourth quartile) promoters from BP directional (on the left), BP divergent (in the middle) and NP (on the right) clusters. Same comparison is shown at the DHS (third line) shared between OSC and S2 cell lines (on the left), uniquely active in OSC (in the middle) and from the first quartile in S2 cell lines (on the right).

#### 4.3.9 Genes closer to each other show low expression plasticity in *D.melanogaster* and vice versa

Particular chromatin and sequence landscapes were observed in the intergenic region between divergent genes, characterized by a strong reduction in GC content and higher DNA-encoded fragility (paragraph 4.3.6). Prompted by this finding, a deeper characterization of the divergent organization followed, since a relationship between sequence and chromatin features with the gene organization was suspected. To systematically investigate that, three main gene organizations were considered based on the TSS-TSS and TSS-TTS distances: divergent, tandem and distant. Tandem cluster included genes that are closer to each other and transcribed in the same direction. Distant cluster included genes that are at least 1Kb far away from any TSS or TTS. To exclude alternatives promoters, only the promoter with the highest CAGE count was considered per gene.

Firstly, promoter shape and expression level were analyzed in each gene organization. Regarding promoter shape, our TSS analysis split genes in around 62% BP and 38% NP (Figure 18.A). Interestingly, a strong enrichment for the BP shape was retrieved in divergent genes, with the divergent\_500 cluster constituted by BP promoters in more than 90% of the cases. Instead, enrichment for the NP shape was found in tandem\_1000 and distant clusters. Regarding expression level (Figure 18.B), divergent genes showed a strong enrichment for the first two quartiles (about 90% in divergent\_500). In parallel, enrichment for low or not expressed genes was obtained in tandem\_1000 and distant clusters. Particularly, the distant\_4000 cluster was constituted by genes of the third and fourth quartiles in more than 75% of the cases. Finally, the expression level among gene organizations was also analyzed based on the promoter shape separation (Figure 18.C). As expected, strong enrichment for high and low expressed genes was found in BP and NP promoters, respectively. However, the divergent organization was still enriched for active genes independently on the promoter shape. In parallel, distant genes were found enriched for inactive genes even with a BP shape. Regarding the tandem organization, no significant enrichment was measured in all these analyses, with the exception of the tandem\_500 cluster, in which slight enrichments for the BP shape and for active genes was observed.

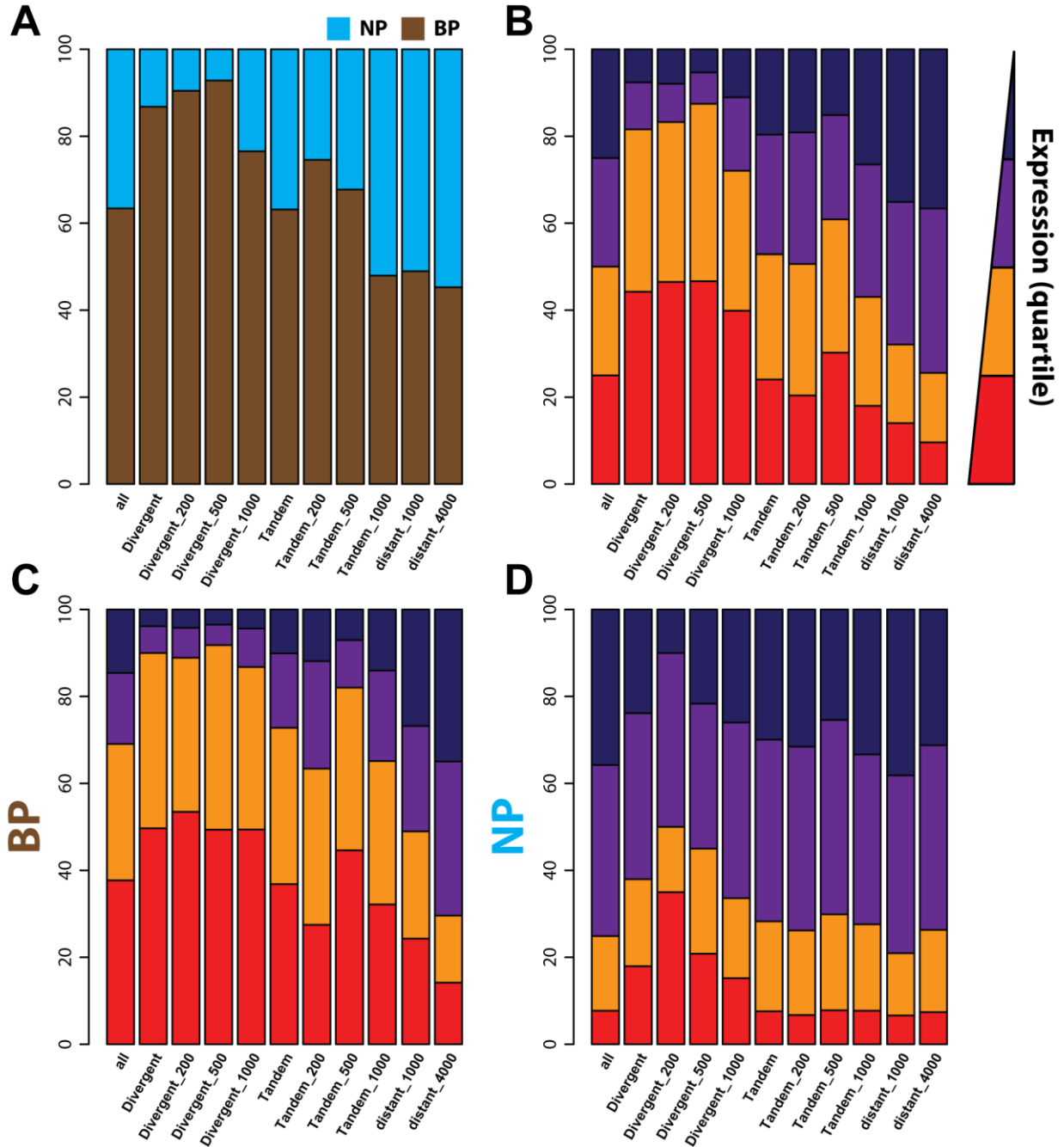
Collectively, these data indicate that genes with low expression plasticity, characterized by a BP promoter shape and high expression in S2 cells, tend to be closer to each other along the



## *Results*

genome. In contrast, genes with high expression plasticity, characterized by a NP promoter shape and low or no expression in S2 cells, are more isolated into the chromatin.

## Figure 18



**Figure 18: Distribution of promoter shape and S2 cell expression level within gene organizations.** (A) Percentage partition of NP and BP promoter shapes in each analyzed gene organization (divergent, tandem and distant). (B) Same partition, but of expression data divided in quartile. Quartiles from blue to red are ordered by increasing expression, so the first quartile (red) contains highly expressed genes. (C) Same partition of (B), but considering only genes with a BP promoter shape. (D) Same partition of (B), but considering only genes with a NP promoter shape. In each panel the all column refers to the entire pool of analyzed genes.

#### 4.3.10 Genes with high and low expression plasticity show different GC profile around promoters

Prompted by the observation that divergent\_500 genes with a NP promoter shape were enriched for high expressed gene and distant\_4000 genes with a BP promoter shape were enriched for genes with low or no expression, a further characterization on these gene organizations was carried out by a GO terms analysis. Only significant terms ( $P < 0.05$ ) with a ratio measured/expected higher than 2 were considered. In parallel, the same analysis was also carried out on the tandem\_500 cluster due to the similarity in distance with the divergent one, but with a different gene organization.

Significant results were observed in all three gene organizations when genes of the first quartile in expression level were considered. Specifically, for BP genes, enrichment for cell cycle regulation and metabolic processes in divergent\_500, and biosynthetic and transporting processes in tandem\_500 were obtained. Strikingly, developmental processes were found enriched in BP distant\_4000 (Figure 19). In parallel, for NP genes, enrichment for regulated processes and developmental terms in distant\_4000, and no significant results in tandem\_500 were retrieved. Strikingly, translational and metabolic processes were found enriched in NP divergent\_500 (data not shown).

To investigate what information were behind the fact that gene organization better captures expression plasticity than promoter shape, DNA and chromatin landscapes were analyzed. Regarding the DNA landscape, the GC content profile was again considered. Interestingly, a remarkable GC depletion was found upstream of the TSS in NP divergent\_500 and tandem\_500 clusters (Figure 19.A & Figure 19.B). In particular, the GC content profile of NP divergent\_500 strongly resembled the BP counterpart. In parallel, the GC content profile of BP distant\_4000 resembled the NP profile counterpart, at least for the first and second quartiles (Figure 19.C). A completely different GC profile was detected for the third and fourth quartiles.

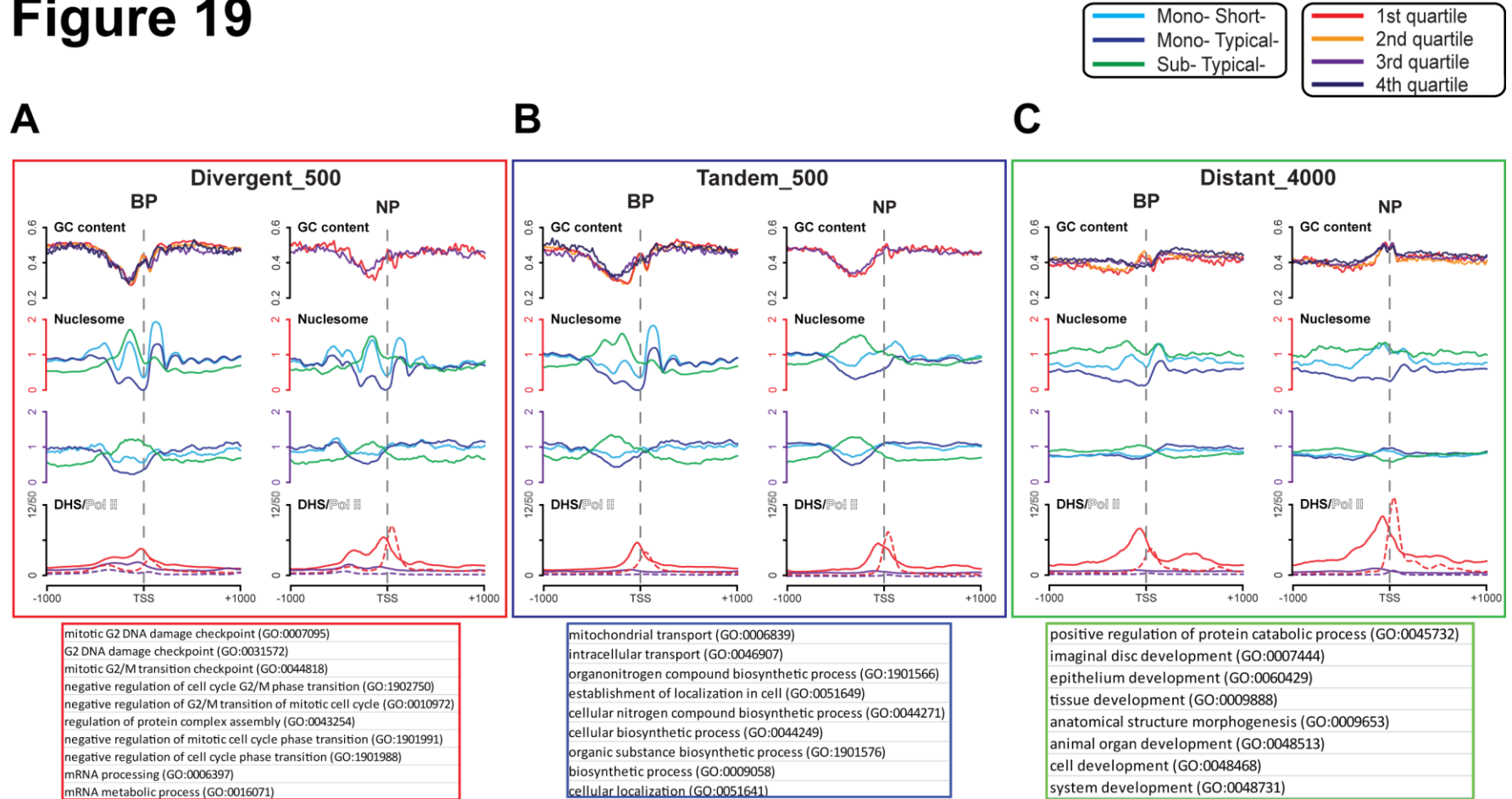
Similar parallelisms were found also at the nucleosome level. Indeed, NP divergent\_500 showed a differential MNase-seq signal similar to the BP counterpart, characterized by a canonical nucleosome pattern and higher and sharper occupancy of mono- short- and sub-typical- at the -1 position. In contrast, both BP and NP distant\_4000 were characterized by a strongly reduced canonical nucleosome pattern and by broader and activity-dependent MNase sensitivity around the TSS. Moreover, these parallelisms were also observed in the DHS profile.

## *Results*

Indeed, higher and broader accessibility upstream of the TSS was measured in BP distant\_4000 than BP divergent\_500, whose signal reassembled the NP counterparts. Interestingly, the paused Pol II profile followed what expected by the promoter shape, namely lower and higher occupancy in BP and NP promoters, respectively.

In conclusion, the fact that NP divergent\_500 is enriched for housekeeping processes and BP distant\_4000 is enriched for developmental ones indicates that other genic features than promoter shape can better correlate with expression plasticity. These results indicate that GC content and chromatin landscapes can be considered as candidates due to higher consistency in distinguishing genes with different expression plasticity.

## Figure 19



**Figure 19: Sequence and chromatin landscapes through gene organization.** Profiles of the GC content (25 bp bins) (first line) and mean coverage of mono- short-, mono- typical-, sub- typical- (second and third lines), DHS signal and Pol II (fourth line) around BP and NP promoters of divergent\_500 (A) tandem\_500 (B) and distant\_4000 (C) genes. The second and third lines are plotted around promoters of the first (red) and third (purple) quartiles, respectively. In the fourth line, DHS and Pol II signals are depicted with unbroken and dashed line, respectively. NP promoters of divergent\_500 genes show similar GC content and chromatin profiles to the BP counterpart. BP promoters of distant\_4000 show similar GC content and chromatin profiles to the NP counterpart. At the bottom, the most enriched GO terms are listed for each cluster in a descending order of the measured/expected ratio (at least higher than 2). Only BP genes from the 1st quartile are reported. GO terms search was performed with the PANTHER (release 2017.04.13) classification using standard parameters. Only p-values < 0.05 are shown.

#### **4.3.11 Sequence and chromatin landscapes are connected in divergent promoters**

The divergent organization was found particularly enriched for gene with low expression plasticity. Moreover, a higher DNA-encoded nucleosome fragility was measured in this gene organization, which was independent on the promoter shape. To better investigate the relationship between sequence and nucleosome landscapes, divergent genes were clustered based on the TSS-TSS distance. Meanwhile, to compare the results with other gene organizations, tandem genes were also analyzed and clustered based on the TSS-TTS distances.

Firstly, TSS-TSS and TSS-TTS distance distributions within the gene organizations were calculated. In regard of divergent genes, the TSS-TSS distance distribution showed a remarkable low density below 200bp and the highest peak around 270bp (Figure 20.B). This finding suggests that divergent genes are often spaced by a distance sufficient to accommodate a shared -1 nucleosome. In contrast, the TSS-TTS distance distribution of tandem genes showed a normal distribution skewed toward short distances, and with a peak around 200 bp (Figure S8.B). This indicates that a -1 nucleosome within the opposite gene is not rare.

Subsequently, BP divergent and tandem genes were binned accordingly to the TSS-TSS and TSS-TTS distances, respectively. In each bin, GC content and nucleosome landscapes were mapped. In regard of the divergent organization, a completely different GC content and chromatin landscapes were measured between bins with a TSS-TSS distance less than 200bp and other ones. Specifically, the GC depletion upstream of the TSS was found reduced and sharper compared to other bins (Figure 20.A). Indeed, an intergenic region less than 200 bp cannot include a -1 nucleosome due to steric hindrance. Therefore, evolution of a sequence landscape that induces nucleosome fragility with a drop in GC content does not favor the transcription of these genes. For distances longer than 200 bp, a sequence landscape characterized by a stronger and broader GC depletion between the two TSSs was observed, which covered the entire intergenic region. Interestingly, this GC landscape is preserved among bins. At the same time, a broadening of the GC depletion was also measured with the increase of the TSS-TSS distance. Strikingly, a parallelism between variations of the GC profile with the nucleosome organization among bins was observed. Specifically, a shifting of the -1 and +1 positions was measured according to the lowest point of the GC depletion and to the asymmetrical GC content, respectively. With distances longer than 450-500 bp, a further broadening of the GC depletion

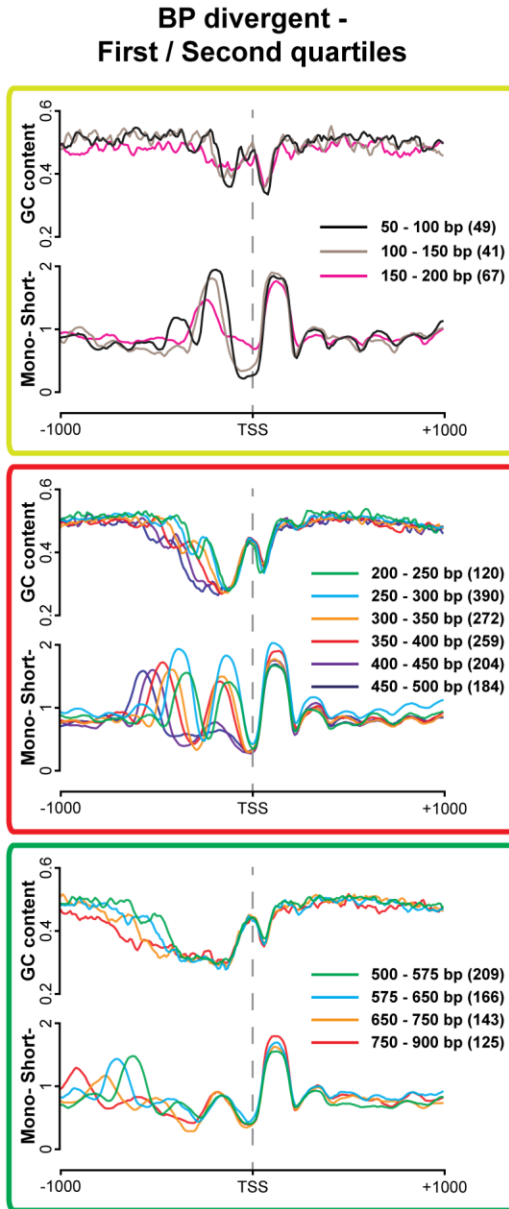
and a presence of two -1 nucleosomes were obtained upstream of the TSS. Indeed, in these cases, the intergenic region is large enough for two nucleosomes to be accommodated. Interestingly, for TSS-TSS distances longer than 450 bp, no significant shift in the -1 position was measured among bins.

In regard of the tandem organization, a GC profile similar to the one found in the BP directional cluster was observed among bins (Figure S8.A). As already described for the divergent organization, also tandem genes showed a broadening of the GC depletion upstream of the TSS with the increasing of the TSS-TTS distance. In tandem\_200 bins, the -1 position overlapped with an asymmetrical GC content, instead of the lowest point of the GC depletion. Indeed, these nucleosomes cannot be considered as typical -1 ones, since they occupy the TTS of, or are located within, the upstream gene. Instead, other tandem bins showed similar nucleosome features previously described for the divergent organization with a TSS-TSS distance longer than 450 bp. In fact, the -1 and +1 positions consistently overlapped with the lowest point of the GC depletion and with an asymmetrical GC content, respectively.

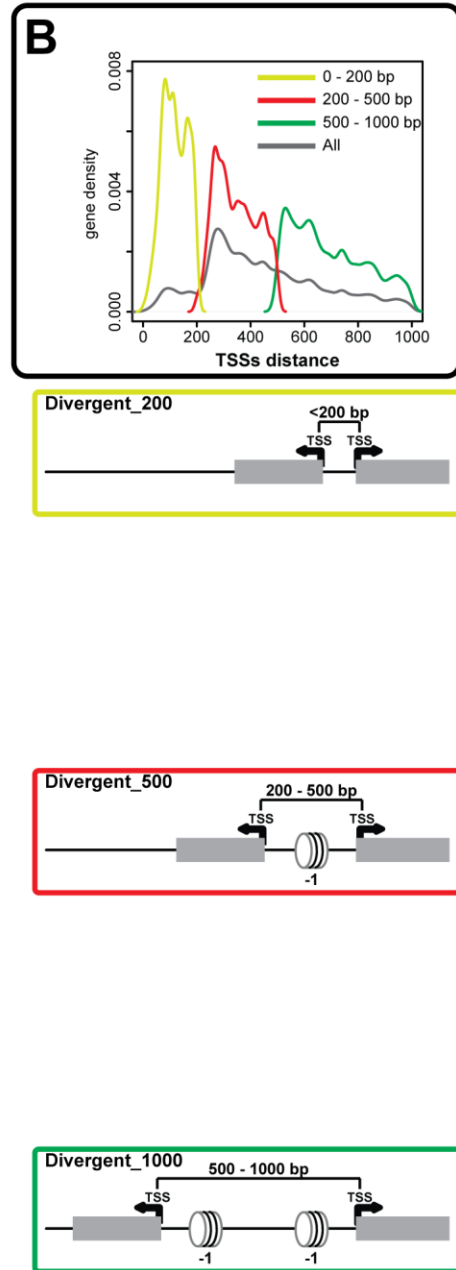
In conclusion, an intimate relationship between GC content and nucleosome positioning was found particularly in the divergent organization. For these genes, data point to an evolution of a sequence landscape that favors a distance of 200-500 bp between TSSs. Indeed, a distance longer than 200 bp can accommodate a shared -1 nucleosome, which is also characterized by DNA-encoded fragility due to the low GC content. Most importantly, the divergent 200-450 bp bins, the -1 and +1 positions are consistently located on the lowest point of the GC depletion and on an asymmetrical GC content, respectively. How DNA and chromatin landscapes of the divergent organization are related to the strong enrichment for genes with low expression plasticity remains to be further elucidated.

# Figure 20

**A**



**B**



**Figure 20: Relationship between TSSs distance with GC content and chromatin profiles in divergent genes.** (A) Profiles of GC content (25 bp bins) (first line of each panel) and mean coverage of mono- short- (second line of each panel) around BP promoters of divergent genes from the first and second quartiles. The legend in each panel shows the TSSs distance binning and the gene count in brackets. Divergent TSSs with a distance between 50 – 200 bp are shown in the upper panel (yellow). Divergent TSSs with a distance between 200 – 500 bp are shown in the middle panel (red). Divergent TSSs with a distance between 500 – 1000 bp are shown in the lower panel (green). (B) Smoothed gene density distribution in function of the TSSs distance among divergent genes. The gray line is the distribution for all divergent genes with a TSSs distance between 0 – 1000 bp. Colored distributions match the distance distinction used in (A).



## 4.4 PART III – MNase-Short-ChIP-seq

For genome-wide measurements of chromatin-related features, ChIP-seq is the method of choice. Commonly, chromatin fragmentation is performed through mechanical, such as sonication (X-ChIP-seq), or enzymatic approaches, such as MNase digestion in a typical-digestion range (MNase-Typical-ChIP-seq). Both methods could deplete the starting material of MNase-sensitive nucleosomes, which are more preserved with a MNase short-digestion. Therefore, in this part, the suitability of both approaches in measuring chromatin-related features on MNase-sensitive nucleosomes was investigated. In parallel, a new and simplified protocol was developed by using MNase short-digestion for chromatin shearing. Specifically, FA cross-linked chromatin was briefly digested with MNase to yield 30-40% of mono- plus sub-nucleosomes compared to the bulk chromatin. MNase digestion protocol was then joined to ChIP by buffer adjustments. Finally, pulled-down mono- and sub-nucleosomes were isolated by size-selected after library preparation. To test this approach, chromatin-related features typically found in active chromatin regions were considered, since these regions are enriched for MNase-sensitive nucleosomes. Hence, four histone tail PTMs strongly enriched around active promoters and enhancers were analyzed: H3K4me3; H3K9ac; H3K18ac; H3K27ac. These histone marks are not specific to *D.melanogaster* (Kharchenko et al., 2011; Yin et al., 2011), but they are also enriched in promoter and enhancers of worm and human (Ernst et al., 2011; Ho et al., 2014).

### 4.4.1 MNase-Short-ChIP-seq: a new approach to better measure chromatin features in active genomic regions

As first step, a validation of MNase-Short-ChIP-seq method and advantages was carried out by comparing DNA features and mapping of the starting material and pull-downs against the other two shearing approaches: MNase-Typical-ChIP-seq and X-ChIP-seq.

To ascertain that MNase-sensitive nucleosomes were indeed better retained in MNase-Short-ChIP-seq, starting materials were compared by measuring di-nucleotide frequencies (Figure S9). Indeed, enrichment for WW and depletion for SS di-nucleotides were observed in the input of MNase-Short-ChIP-seq, exclusively. In contrast, almost no changes in respect to the genomic average were measured in the inputs of MNase-Typical-ChIP-seq and X-ChIP-seq. Sequence features from pulled-down fragments were then analyzed. Interestingly, in MNase

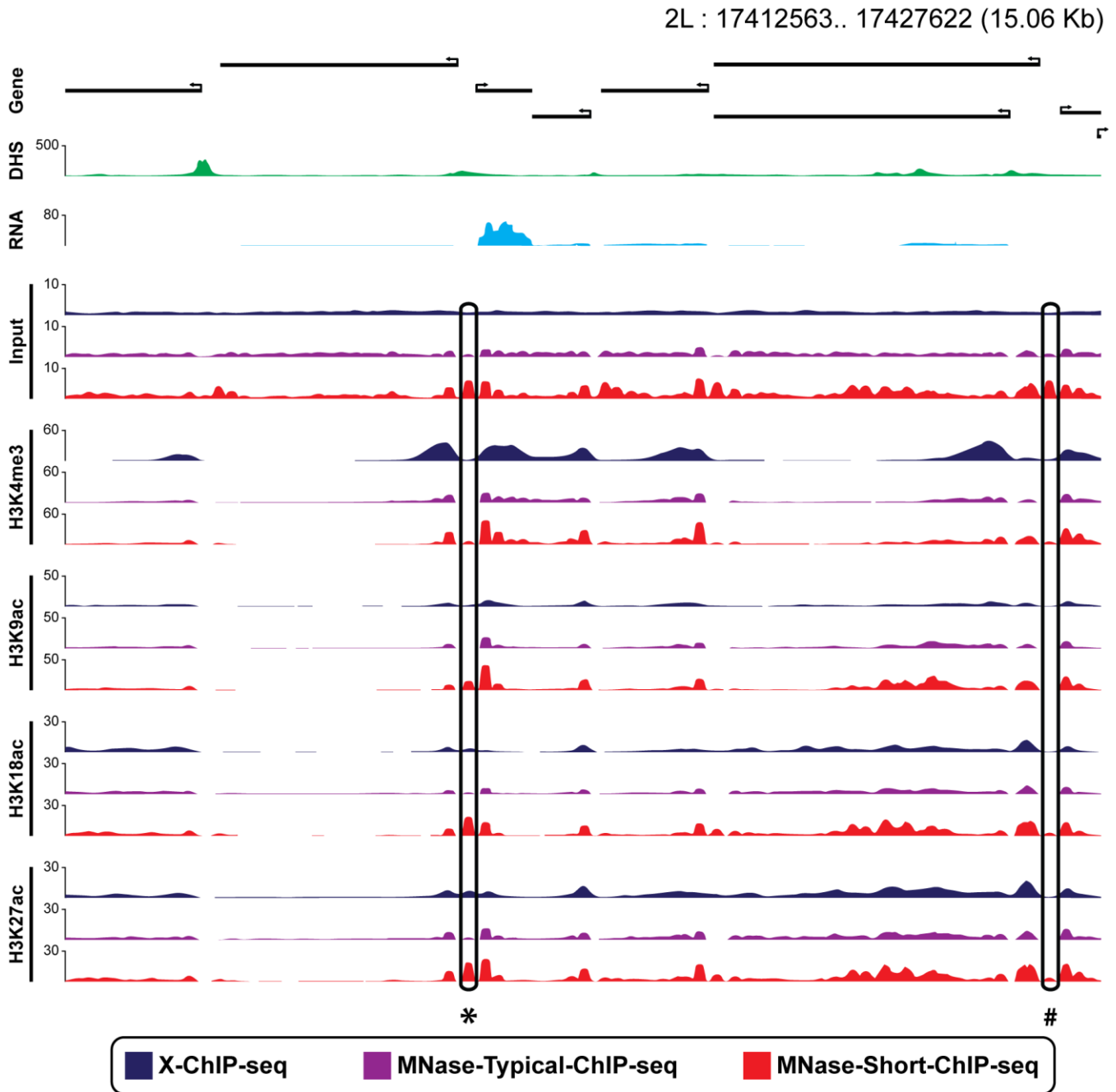
short-digested chromatin, H3K18ac and H3K27ac showed a similar profile to the starting material. In contrast, H3K9ac, and more dramatically H3K4me3, showed a decrease in WW and an increase in SS frequencies. Remarkably, di-nucleotide frequencies from pull-downs obtained with the other two shearing approaches were completely different, in which a general depletion of WW di-nucleotide was observed.

In conclusion, MNase-Short-ChIP-seq improves the retention of MNase-sensitive nucleosome both in the starting material and pull-downs. Moreover, histone mark pull-downs are characterized by different DNA features, which can derive from a differential occupancy among nucleosome positions.

Subsequently, input and histone tail pull-downs from the three shearing approaches were mapped on the genome. Hence, MNase-Short-ChIP-seq was validated and its advantages were evaluated genome-wide through comparison with the other two shearing approaches (Figure 21). Firstly, a higher signal to noise was found in MNase-Short-ChIP-seq, mostly biased in active chromatin regions, as indicated by the proximity of the input signal to DHSs. Secondly, a nucleosome resolution of the data was obtained only within MNase digested samples. Thirdly, a great overlapping was reached among histone mark profiles from different shearing approaches. Finally, differential occupancy among histone marks was better captured by MNase-Short-ChIP-seq, and it was often measured on the -1 nucleosome. For instance, in the reported genome-wide screenshot, enrichment for acetylation marks, but not for H3K4me3, was measured on the asterisk position in MNase-Short-ChIP-seq. Instead, this differential occupancy was not detected in MNase-Typical-ChIP-seq, most likely due to the digestion level bias, and it was mitigated in X-ChIP-seq, most likely due to lower signal to noise ratio and resolution of the data.

Moreover, the specificity of the method was also evaluated. As the asterisk position, also the hash mark position corresponded to a -1 nucleosome. Furthermore, a divergent organization and a similar occupancy in the starting material were also shared between the two positions. Meanwhile, no signal from the histone marks was obtained on the hash mark position in MNase-Short-ChIP-seq, most likely due to the inactivity of the closer genes.

# Figure 21



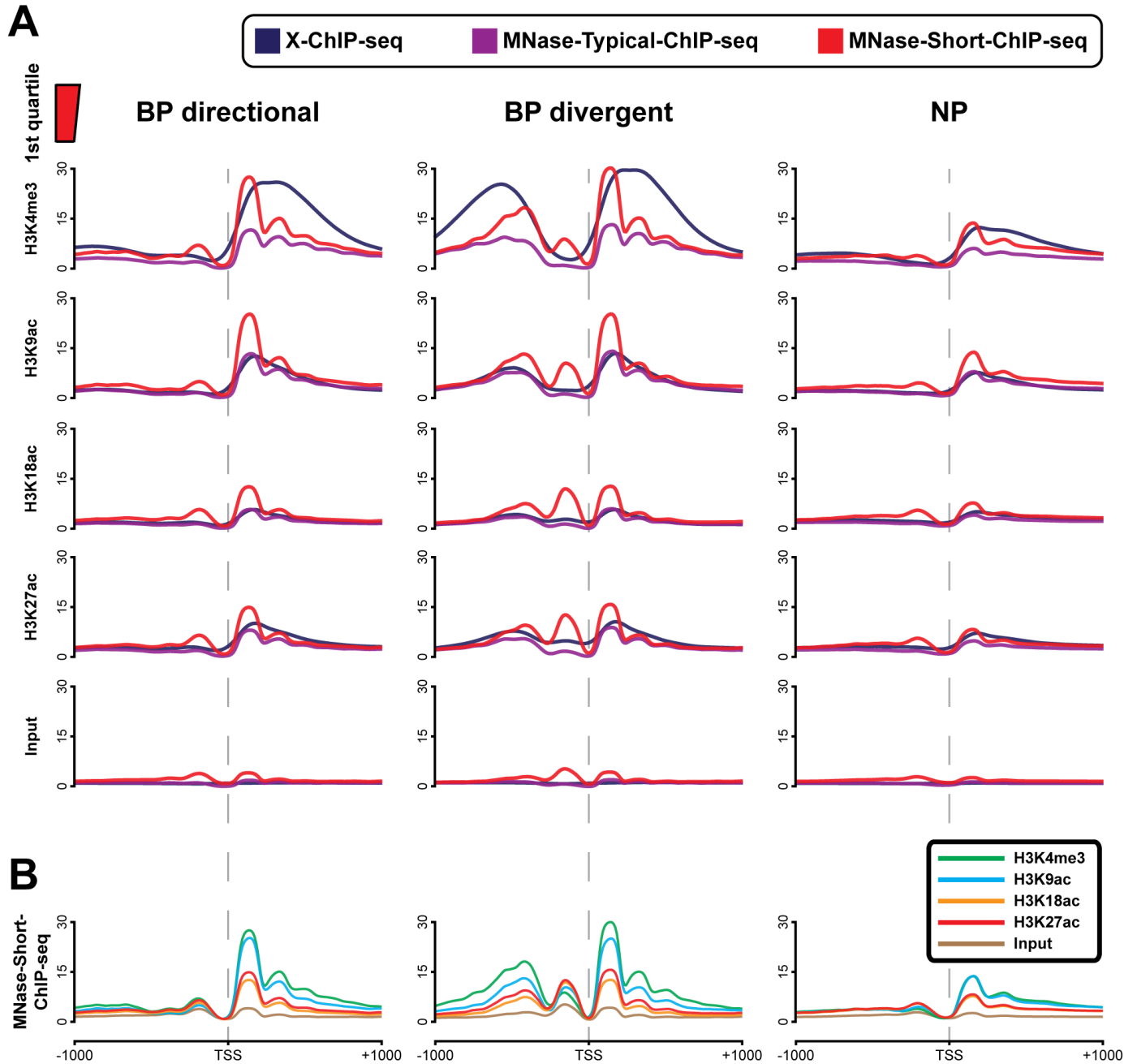
**Figure 21: Genome-wide screenshot of histone tail PTMs pull-downs differently sheared.** A screenshot showing histone tail PTMs pull-down profiles in a window of around 15 Kb of the chromosome 2L. From the top, it is shown gene localization, DHS and RNA-seq signals. Inputs and pull-downs from the three shearing approaches are shown further below. For a better evaluation different scales are applied among tracks. Asterisk and hash mark positions are highlighted by black boxes. Both nucleosomes correspond to a -1 position of divergent genes. They have similar occupancy in the MNase-ChIP-Short-seq input, but differential enrichment for histone tail PTMs. Indeed, the asterisk nucleosome is located upstream of a highly expressed gene and vice versa for the hash mark nucleosome.

Parallelisms and differences among the three shearing approaches were also confirmed by mapping the data around promoters clustered on their shape. In regard of differences, a higher occupancy for H3K4me3 on the -1 position and for acetylation marks on +1 and -1 positions were obtained in the MNase-Short-ChIP-seq (Figure 22.A).

Successive to the validation, MNase-Short-ChIP-seq tracks were mapped around promoters, clustered by their shape, to measure differential occupancy among histone marks (Figure 22.B). At the -1 position, in BP directional promoters, similar occupancy between acetylation and H3K4me3 was observed. In parallel, BP divergent and NP promoters showed enrichment for H3K18ac and H3K27ac over the methylation mark. At the +1 position, a strong enrichment for H3K4me3 and H3K9ac over the other two acetylation marks was measured in all promoter clusters. Interestingly, in NP promoters, a higher occupancy of H3K9ac on the +1 nucleosome and a broader distribution of all acetylation marks in both directions around the TSS were obtained. This observation most likely derives by the close proximity with active enhancers, typically found around developmental genes.

Here, a new approach to better measure chromatin-related features in active chromatin regions was developed and validated. By preserving MNase-sensitive nucleosomes, MNase-Short-ChIP-seq reveals a higher, specific and differential occupancy of H3K4me3, H3K9ac, H3K18ac and H3K27ac around promoters, particularly at the -1 and +1 positions.

# Figure 22



**Figure 22: Histone tail PTMs profiles around promoter clusters and technical comparison among shearing approaches.** (A) Mean coverage of input and histone tail PTMs pull-downs from the three shearing approaches around BP directional (on the left), BP divergent (in the middle) and NP (on the right) promoter clusters. Only genes from the first quartile (high expression) are considered. (B) Same data of (A), but taking in consideration only MNase-ChIP-Short-seq data in order to better compare differential histone tail PTMs enrichment among promoter clusters.

#### **4.4.2 Histone mark occupancy is influenced by activity and distance of closer genes**

In part II a relationship between expression plasticity and gene organization was found. Moreover, sequence components were considered within this relation by analyzing the GC content of gene organizations. Here, activity components were considered by using histone mark profiles obtained by MNase-Short-ChIP-seq. In detail, a relationship between histone marks occupancy and gene organization was investigated by considering activity and distance of closer genes.

The influence played by the activity of closer genes on the histone mark deposition was analyzed for the divergent and tandem organizations. Indeed, active divergent\_500 and tandem\_500 genes were divided in two clusters according to the expression level of the closer gene (Figure 23). The high vs high cluster included divergent or tandem genes in which the closer gene is also active. Instead, the closer gene is inactive in the high vs low cluster. Finally, to ascertain that results derived uniquely from the activity of closer genes, other variables, such as promoter shape and expression level, were removed. Hence, histone marks were exclusively mapped around genes with a BP promoter shape and with a similar average expression level ( $3 \leq \log_2(\text{FPKM}) \leq 8$ ). To confirm the correctness of this approach, sequence and activity components were compared between the two clusters. Indeed, a similar GC content, chromatin accessibility and Pol II profiles around the TSS were obtained between high vs high and high vs low clusters. In parallel, the high vs low cluster showed a reduction of histone marks and Pol II occupancies, and a reduced chromatin accessibility over the opposite gene body.

Strikingly, at the -1 position of divergent genes, a strong dependency on the activity of closer genes was measured for the deposition of all histone marks. In parallel, at the +1 position, H3K9ac also showed similar dependency (Figure 23.A). Moreover, these results were not detectable from MNase-Typical-ChIP-seq, and they were strongly reduced in X-ChIP-seq. Similar results were obtained in the tandem organization. Indeed, occupancy of all histone marks at the -1 position, and occupancy of H3K9ac at the +1 position were higher in the high vs high cluster. However, differences between the two cluster were slighter compared to the divergent organization, most likely due to a reduced nucleosome positioning (Figure 23.B).

The influence played by the distance of closer genes on the histone mark deposition was assessed from the tandem, binned according to the TSS-TTS distance, and distant organizations.

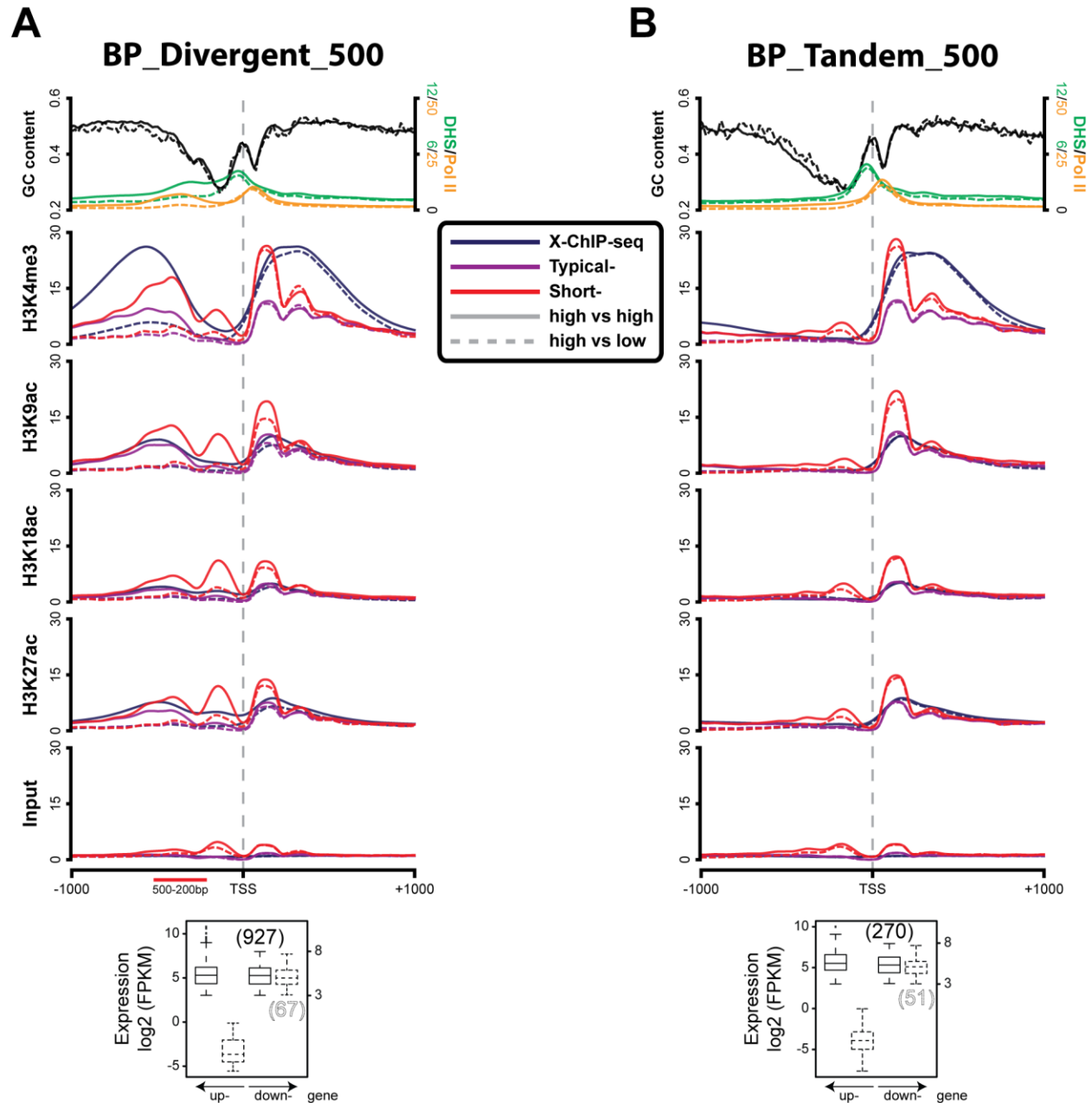
As performed for the activity dependency, only genes with a BP promoter shape and a similar average expression level ( $3 \leq \log_2(\text{FPKM}) \leq 8$ ) were considered (Figure S10).

Interestingly, a shift in the ratio between methylation and acetylation occupancy at the -1 position was observed with the increase of the intergenic space between genes. Indeed, higher H3K4me3 occupancy was measured when the -1 nucleosome is positioned within the gene body of the closer gene. At the same time, a progressive reduction of H3K4me3 and an increase in H3K18ac and H3K27ac occupancies were measured at the -1 position with the increase of the intergenic space. The latter observation was retained in the distant organizations.

Moreover, histone mark profiles of BP distant\_4000 promoters showed features that can be assigned to genes with high expression plasticity. Indeed, a widespread acetylation marks deposition upstream of the TSS and downstream of the +2 nucleosome was observed. Furthermore, BP distant\_4000 showed a stronger enrichment of H3K9ac at the +1 position. As mentioned in the paragraph above, both observations were obtained in NP promoters. Therefore, the fact that distant\_4000 organization is enriched for gene with high expression plasticity, as assessed in PART II of this dissertation, found further confirmation from the histone tail PTM mapping.

Collectively, these results suggest that histone tail PTMs deposition around promoters is influenced by the activity and distance of closer genes, in particular at the -1 position. Indeed, histone mark deposition is higher when divergent or tandem genes are both active. In parallel, acetylation deposition is broader and higher than H3K4me3 when genes are more isolated into the chromatin. Interestingly, these evidence can be detected exclusively by a MNase-Short-ChIP-seq approach, indicating the goodness of this method in capturing novel chromatin features in active genomic regions.

# Figure 23



**Figure 23: Activity context in the histone tail PTMs deposition around the TSS.** Profiles of GC content (25 bp bins) and mean coverage for DHS signal and Pol II are shown in the first line. The same for input and histone mark pull-downs are shown in the other lines. All these profile are measured around BP promoters of genes with a divergent\_500 (A) and tandem\_500 (B) organizations. Only genes with an expression level of  $\log_2(\text{FPKM}) \geq 3$  &  $\leq 8$  are considered. For each gene organization, genes are divided in two clusters accordingly to the expression level of the closer genes. In high vs high (unbroken lines) closer genes are active. In high vs low (dashed lines) closer genes are low or not expressed. Finally, profiles from each shearing approach are shown. At the bottom of each panel boxplots of  $\log_2(\text{FPKM})$  are attached. They indicate the expression level of genes upstream and downstream of the TSS. Unbroken and dashed lines still indicate high vs high and high vs low clusters.



# 5 DISCUSSION

## 5.1 MNase biases

### 5.1.1 MNase sequence bias does not determine nucleosome mapping

MNase-seq is still the method of choice for nucleosome mapping. Nevertheless, it is affected by a strong preference of cutting with A and T at the 5'. This study characterized the MNase sequence bias at the cut site on both genomic naked DNA (gDNA) and chromatin in great detail and in a genome-wide fashion. Our data reveals that MNase strongly prefers to cut within a TA di-nucleotide. However, the surrounding sequence content around the cut site is complex. Therefore, it is important to distinguish the preference of cutting within TA and AT-rich sequence, since only the first was captured in this study. Similar sequence complexity around the cut site was previously observed, but only on a limited pool of DNA sequences (Dingwall et al., 1981; Horz and Altenburger, 1981).

Most importantly, a reduction of the sequence bias was observed in the digestion of the chromatin. This is most likely due to the nucleosome occlusion on most of the genome, whose protection is the result of MNase digestion.

This study also asked how MNase sequence bias maps around CREs of *D.melanogaster*. Some degree of correlation among sequence bias, nucleosome mapping and GC content was observed, as previously reported in yeast (Chung et al., 2010), these correlations can mask important details that can be retrieved through comparison of gDNA and nucleosomal tracks on specific genomic regions. Our results measure that nucleosome map is not determined by the sequence bias, at least around promoters. At the same time, higher overlap between gDNA and nucleosomes was obtained around DHSs and TTSSs. Moreover, the nucleosome organization observed around these CREs agrees with the chemical approach in yeast, which is not affected by the TA sequence bias (Brogaard et al., 2012).

Taken together, this study represents the most comprehensive characterization of the sequence bias around the cuts site and it rejects the hypothesis that MNase sequence bias determines how nucleosomes are mapped around promoters.

### **5.1.2 Evidence of MNase digestion within the nucleosome driven by nucleosome unwrapping**

MNase is also able to cut within the nucleosome. This study characterized the intra-nucleosomal digestion with unprecedented detail through prolonged digestion level on native and FA cross-linked chromatin. We observed that MNase digestion within the nucleosome is not only a continuous process, but mostly proceeds in a discrete and asymmetric fashion.

The intra-nucleosome digestion was observed as discrete by characterizing the sub-nucleosome fragment distribution, in which four peaks of 128, 105, 90 and 75 bp emerged from a continuous distribution. The same pattern was observed in other studies conducted in different species, such as human and mice (Hasson et al., 2013; Ishii et al., 2015). Taking into account asymmetric digestion, MNase cuts within the nucleosome well overlap with regions of low strength in the DNA–histone interactions (Hall et al., 2009). Therefore, our results indicate that MNase digestion within the nucleosome is mostly driven by the DNA–histone binding strength and does not simply digest the nucleosomal DNA in a continuous way.

The asymmetry of the intra-nucleosome digestion was demonstrated by mapping sub-nucleosome fragment centers around called nucleosome positions in a so called V-plot (Henikoff et al., 2011). Although MNase intra-nucleosome digestion on bound DNA cannot be totally excluded, this study suggests that nucleosome unwrapping, either driven by spontaneous mechanisms or externally induced, mainly determines MNase digestion of nucleosomal DNA. Most importantly, this result is in strong agreement with a recent biophysical study, in which asymmetric nucleosome unwrapping was observed in reconstituted nucleosomes with the Widom 601 sequence (Ngo et al., 2015). Moreover, Ngo and colleagues revealed that DNA sequence, specifically the TA 10 bp periodicity, has a major role to determine from which side nucleosomes undergo to unwrapping. In parallel, we found a relationship between MNase asymmetric intra-nucleosome digestion and underlying DNA sequence, specifically the GC content. The parallelism between asymmetric unwrapping and asymmetric MNase digestion suggests a key experiment where MNase is used as a probe to test all DNA-histone interactions on the Widom 601 nucleosome. To the best of our knowledge, such an experiment has not been performed. However, in the attempt to combine MNase and exonuclease III digestions for nucleosome mapping, Nikitina and colleagues characterized MNase cuts just in close proximity at the entry/exit sites of reconstituted nucleosomes with the Widom 601 sequence (Nikitina et al.,

2013). Strikingly, they found higher MNase cut frequency on the nucleosome side that is more prone to asymmetric unwrapping, as subsequently demonstrated by Ngo and colleagues.

In conclusion, this study reports asymmetric MNase intra-nucleosome digestion in higher eukaryotes and in a genome-wide fashion for the first time. Moreover, our results show that MNase digestion within the nucleosome is mostly driven by biophysical properties of the DNA-histone interactions. Indeed, the observed discrete and asymmetric intra-nucleosome digestion strongly agrees with DNA-histone binding properties assessed by published biophysical studies.

### 5.1.3 The +1 nucleosome is asymmetrically digested by MNase

In BP promoters, the +1 nucleosome is positioned on a GC asymmetric sequence with the half proximal to the TSS having a lower GC content than the distal half. Following our lines of investigation, an asymmetric MNase digestion was found within this nucleosome position by favoring the side with lower GC content.

In this study, asymmetric MNase digestion within the +1 nucleosome was demonstrated through characterization of sub-nucleosome cuts, whose frequency strongly decreased toward the nucleosome dyad in inactive genes. Therefore, this result strongly agrees with the notion that DNA-histone interactions are stronger close by the nucleosome dyad, as suggested by mechanical nucleosome unzipping (Hall et al., 2009) and by DNA mutations on the dyad nucleosomal DNA (Bintu et al., 2012).

The asymmetrical GC content underlying the +1 nucleosome is not a unique feature of *D.melanogaster*. A similar scenario was also observed in *S.cerevisiae* and *S.pombe*, even if it is less pronounced in the latter (Moyle-Heyrman et al., 2013). In humans, some gene clusters were characterized with a +1 nucleosome within an asymmetrical CpG island (CGI) content, whose occurrence was correlated with higher nucleosome depletion (Fenouil et al., 2012). Moreover, asymmetric protection of the +1 nucleosome was also showed through chemical mapping in budding yeast (Ramachandran et al., 2015), and correlated to activity components. In contrast, our data indicate that also the DNA sequence plays a role, since asymmetric protection was even found in inactive genes.

Interestingly, asymmetric GC content within the nucleosome was also observed in the di-nucleosome fragments, in which higher GC content was measured in nucleosome halves surrounding the uncut linker DNA. This could imply a role of nucleosome unwrapping in

determining differential MNase accessibility between linker DNAs. In this regard, a deeper investigation on how asymmetric unwrapping could influence higher order structures of the chromatin is necessary. Some attempts in pulling chromatin fibers apart were made, finding that nucleosome embedded in fibers were more stable (Meng et al., 2015). However, Meng and colleagues could not resolve the transition from chromatin fibers to single nucleosomes perhaps due to spontaneous unwrapping at the entry/exit of the nucleosome, as the authors suggested.

In conclusion, our results support the idea of asymmetric unwrapping involving the proximal half to the TSS of the +1 nucleosome in BP genes. It can support transcription directionality, as previously suggested (Ngo et al., 2015), and/or facilitate the overcoming of the +1 nucleosome barrier. In regard of the latter, the +1 nucleosome was shown to form a barrier against transcription at least three times stronger than downstream nucleosomes (Teves et al., 2014b). Moreover, this model can be also energetically favored since spontaneous unwrapping was found energetically inexpensive at the entry/exit of the nucleosome (Koopmans et al., 2009). Finally, this idea is also supported by biophysical evidences. Indeed, asymmetric unwrapping can induce the octamer-to-hexasome transition by forming a DNA tear-drop structure that could expose the proximal H2A-H2B dimer for release (Chen et al., 2017), thus favoring nucleosome disassembling.

## **5.2 Differential MNase-seq**

### **5.2.1 Differential MNase-seq: a powerful tool to study the nucleosome organization**

Differential MNase-seq has enormous advantages over a single digestion level to study the nucleosome landscape from a qualitative point of view. Indeed, differential MNase-seq can be used to assess chromatin properties by measuring nucleosome MNase sensitivity. Some nucleosomes, such as the -1 position upstream of the TSS and nucleosomes within DHSs are characterized by strong MNase sensitivity. In contrast, other nucleosomes are characterized by being insensitive to MNase digestion.

This study showed that several nucleosome populations can be detected by using a differential MNase-seq. Moreover, we also demonstrated that nucleosome populations undergo transitions from longer to shorter fractions with different timing throughout a differential

MNase-seq. Therefore, each nucleosomal fraction in each digestion level contains a unique proportion of nucleosome populations. Most importantly, transitions among fractions can be used to assess chromatin features. By analyzing oligo-, mono- and sub-nucleosome separately, our study showed that the transition from oligo-nucleosomes to shorter fractions provide more information regarding the primary structure of the chromatin, measured as local chromatin accessibility to MNase. In contrast, the transition from mono- to sub-nucleosomes provides more information regarding DNA-histone binding strength in each individual nucleosome, measured as nucleosome accessibility to MNase.

Furthermore, this study ascertained that nucleosome population content of a sample can be distinguished by the underlying DNA features, such as 10 bp di-nucleotide periodicity and GC content. Specifically, shortly digested samples were characterized by a less pronounced 10 bp periodicity and low GC content, vice versa for samples from prolonged digestion levels. Indeed, two distinct chromatin preparations (native and cross-linked chromatin) of similar digestion level, for which a high genome-wide correlation and high overlapping around CREs was found, showed almost identical DNA features. Hence, DNA features can be used as a metric to assess the digestion level of a sample. Therefore, we propose the DNA comparison as approach to evaluate and standardize samples among different laboratories, as alternative to other approaches in which a 100% digestion of the chromatin in mono-nucleosome is instead required (Rizzo et al., 2012).

In conclusion, differential MNase-seq is a powerful tool to study chromatin features by characterizing nucleosome populations. Most importantly, different kind of information can be retrieved from different fractions and transitions.

### **5.2.2 Differential MNase-seq: a new one-digestion level method**

In this study, a new differential MNase-seq method was developed to distinguish nucleosome populations. Since the transitions from oligo- to shorter fractions is mostly driven by MNase cuts on the linker DNA, consequently better capturing local chromatin accessibility, we decided to focus on the mono- to sub-nucleosome transition. By considering this transition, three nucleosome populations were distinguished: fragile, resistant and average. Fragile and resistant nucleosomes showed enrichment for DNA sequences that disfavor and favor nucleosome wrapping, respectively. Thus, the mono- to sub-nucleosome transition can measure nucleosome

accessibility to MNase by probing the DNA-histone interactions strength. Obviously, also activity components, such as histone variants, transcriptional machinery, chromatin remodelers and TFs can play a role. Indeed, enrichment for published H3.3 and H2A.v pull-downs, mainly from low salt chromatin fractions, (Henikoff et al., 2009) was detected on fragile positions. Similar histone variant enrichment on fragile nucleosomes, specifically H2A.Z, was also found in budding yeast (Xi et al., 2011).

Other studies applied a differential MNase-seq to study MNase sensitivity in *D.melanogaster* (Chereji et al., 2016) or to study chromatin accessibility in a plethora of species (Mieczkowski et al., 2016). Chereji and colleagues used the oligo- to mono-nucleosome transition, Mieczkowski and colleagues used horizontal transitions among digestion levels without fraction separation. None of these studies considered the mono- to sub-nucleosome transition. Very importantly, our differential MNase-seq method in distinguishing nucleosome populations presents three main differences or advantages compared to the published ones.

Firstly, our nucleosome population distinction is less affected by the local chromatin accessibility. Indeed, computing the other two abovementioned transitions with our data (oligos and monos scores), MNase sensitivity was not detected around inactive CREs, since both scores mainly captured chromatin accessibility instead of probing nucleosome features. In stark contrast, our approach was able to discriminate MNase sensitivity in any chromatin context precisely because the nucleosomal DNA-histone binding strength was assessed. Furthermore, oligos and monos transitions are based on the linker DNA digestion, which is strongly affected by the MNase TA preference of cutting, as shown in this study. Moreover, applying a differential MNase-seq to study chromatin accessibility can be confining, since other well-established and simpler methods are available for this purpose, such as DNase-seq, FAIRE-seq and ATAC-seq (Tsompana and Buck, 2014).

Secondly, our method was performed on called nucleosome positions, instead of genome binning. This increases the probability that the considered sub-nucleosome fragments derive from nucleosome particles. Although a partial contribution from non-histone protections cannot be excluded, genomic binning cannot distinguish the source of protection within each bin at all. Thus, it affects nucleosome mapping, detecting mainly unspecific chromatin accessibility.

Thirdly and most importantly, our method is based on the mono- to sub-nucleosome transition from only one digestion level, which is a faster setup, since a great robustness against

the digestion level bias was measured. Thus, it comprises the positive aspect of a differential MNase-seq without the tedious practice of carrying out multiple digestion levels.

Taken together, we developed a new differential MNase-seq method to distinguish nucleosome populations based on the mono- to sub-nucleosome transition within one digestion level. Therefore, our method focuses on MNase nucleosome accessibility by capturing the DNA-histone binding strength.

### 5.2.3 Functional plasticity correlates with a dual nucleosome organization

Our differential MNase-seq method was applied to study nucleosome organization around promoters, enhancers and TTSs. Moreover, nucleosome fragility and resistance were evaluated in respect to sequence and activity components. Sequence components were considered by overlapping fragile and resistant nucleosomes with the underlying GC content. Activity components were considered by dividing genes in quartile of expression level and enhancers in quartile of DHS signal strength. Interestingly, a dualism in fragility and resistance landscapes in respect to sequence and activity components was observed among CREs. Indeed, genes with a BP promoter shape were characterized by a nucleosome fragility carved in the DNA sequence upstream of the TSS. In particular, BP genes with a divergent organization showed a stronger sequence driven fragility at the -1 position. In contrast, genes with a NP promoter shape and enhancer showed a DNA-encoded resistance on their CREs and an activity dependent nucleosome fragility. To explain this dualism, functional plasticity can be taken into account. BP promoters are mostly found in housekeeping genes, which are characterized by low expression plasticity and low transcriptional regulation. Therefore, the DNA-encoded fragility upstream of the TSS could support the recruitment of the transcription machinery and boost the basal transcription. On the other hand, NP promoters are mostly found in developmental genes, which are characterized by high expression plasticity and high transcriptional regulation. Therefore, the nucleosome resistance encoded on their TSSs as well as on enhancers could strongly reduce the spurious activation of these elements, which could lead to dangerous consequences to the organism. Thus, the nucleosome fragility observed in NP promoters and enhancers is strictly regulated by active mechanisms.

The divergence in nucleosome organization around promoters was previously detected between genes with low and high expression plasticity in budding yeast, *D.melanogaster* and

other species (Field et al., 2008; Rach et al., 2011; Tirosh and Barkai, 2008). However, nucleosome populations and their relations with sequence and activity components have not been considered. In budding yeast, this dualism was also correlated to sequence features. Indeed, constitutive promoters were characterized by open chromatin and a NDR induced by poly(dA:dT) traits, whereas inducible promoters were characterized by higher nucleosome occupancy and enriched for TATA-boxes (Cairns, 2009). On the other side, it seems that the relationship between DNA and nucleosome is not universal (Moyle-Heyrman et al., 2013) and active mechanisms have evolved according to the genomic features (Li and Du, 2014). As examples, *Tetrahymena thermophila* has an AT-rich genome (22% GC), and local GC bumps drive nucleosome positioning around promoters (Beh et al., 2015). Conversely, mammalian promoters of house-keeping genes are enriched for CpG islands (more than 50% GC) that dictate nucleosome depletion (Fenouil et al., 2012). Noteworthy, in the same study, an anti-correlation was measured between the +1 nucleosome occupancy and GC content in mouse, but only above a GC threshold of 0.58-0.6. Therefore, the correlation between GC content and nucleosome occupancy is not absolute, but rather ranges around certain GC values.

In our study, we found a GC peak within DHSs surrounded by slight GC depletions. This sequence landscape, which is similar to NP promoters, favors the presence of resistant nucleosomes on the enhancers. Indeed, a high intrinsic nucleosome barrier was predicted on these elements (Gaffney et al., 2012). Therefore, enhancer activation strongly relies on TFs and chromatin remodelers to open the chromatin locally. At the same time, the GC depletions surrounding the DHS peak leads to a fragile environment, priming local chromatin accessibility on inactive enhancers. In a parallel project, we used ecdysone stimulation in S2 cells to measure fragility and resistance on responsive DHS peaks. We found out that, transitions from resistant to fragile nucleosomes or vice versa occur, and are most likely driven by activity components, such as pioneer factors and chromatin remodelers.

Downstream of the TTS, higher GC content in BP compared to NP genes was observed, which also overlapped with higher frequency of resistant nucleosomes. This finding indicates the possibility of distinct strategies of transcription termination between promoter shapes. Indeed, chromatin plays a role in the Pol II pausing on the TTS by influencing its dwell time over the polyadenylation signal (Proudfoot, 2016), which is indeed supported by a higher frequency of the G nucleotide (Skourti-Stathaki and Proudfoot, 2014). Higher transcription termination



efficiency was demonstrated able to improve the transcription initiation of the same gene (Mapendano et al., 2010), a strategy that can boost the basal expression of BP genes.

Taken together, nucleosome populations mapping through our differential MNase-seq method detects novel features of the nucleosome organization around CREs. In particular, the dualism between CREs with high and low functional plasticity is mirrored in differences in fragility and resistance landscapes, in which sequence and activity components interplay.

#### **5.2.4 Divergent promoters: a special gene organization**

As mentioned in the previous paragraph, higher DNA-encoded fragility was found in the intergenic region between two divergent protein-coding genes with a distance between TSSs of 200-500 bp. This space permits a -1 fragile nucleosome to fit in. When divergent genes were clustered in distance bins, the shared -1 nucleosome position and the deepest point of the GC depletion upstream of the TSS are consistently co-localized, and were located exactly in the middle between the two TSSs. This sequence landscape could have evolved in order to symmetrically cover the intergenic region with a fragile nucleosome. Moreover, histone tail PTMs deposition on the shared -1 nucleosome showed a strong dependency on the total activity of both divergent genes. Particularly, this was found for the acetylation marks that can weaken inter-nucleosomal interactions and increase the local chromatin accessibility (Widlund et al., 2000). At the same time, all these findings were mitigated or not found in the tandem organization, a gene organization in which two protein coding genes transcribed in the same direction are in close proximity.

We propose that divergent organization with a shared -1 nucleosome has evolved to boost basal expression of both genes through a DNA-encoded fragility, thus facilitating PIC recruitment on both TSSs. In fact, a strong enrichment for the BP promoter shape and for genes with low expression plasticity was found in the divergent organization. Moreover, the divergent organization is evolutionary conserved and its frequency is higher than expected by gene density, even in the longer human genome (Adachi and Lieber, 2002). Furthermore, human divergent genes also show low level of expression plasticity (Lin et al., 2007). Finally, no significant results were obtained by performing similar analysis on the tandem gene organization (Trinklein et al., 2004), which is in accordance with our results. Interestingly, in these studies, divergent genes were shown to be significantly enriched in GO terms for DNA repair and RNA-helicases

(Adachi and Lieber, 2002; Trinklein et al., 2004). From our data, the first quartile of the BP divergent\_500 cluster also shows similar, if not identical, GO enrichments, such as DNA repair, regulation of the cell cycle, RNA processing and other housekeeping processes.

In conclusion, nucleosome population mapping revealed a unique DNA-encoded fragility between divergent genes, and further analyses expanded this finding by supporting the idea that such nucleosome organization is correlated with genes with low expression plasticity and a BP promoter shape.

### **5.2.5 Gene expression plasticity is mirrored in the underlying GC content**

Housekeeping genes are mostly characterized by a BP promoter shape and developmental genes by a NP promoter shape in *D.melanogaster* (Hoskins et al., 2011; Rach et al., 2009). Here, we showed that genes with different expression plasticity correlate stronger with the underlying GC profile than promoter shape. Indeed, when a different criterion was used for gene clustering, namely gene organization, the promoter shape dichotomy with expression plasticity was partially lost. Specifically, NP genes with a divergent organization were significantly enriched for housekeeping GO terms. In parallel, BP isolated genes (distant\_4000) were significantly enriched for developmental processes in both first and second quartiles of expression level. In contrast, the GC content profile still captures the expression plasticity dichotomy, since NP divergent genes showed similar GC content to the BP counterpart and vice versa for BP genes of the distant\_4000 cluster. Interestingly, no significant GO term enrichment was found for BP distant\_4000 genes that are low or not expressed in S2 cells (data not shown). Indeed, these genes showed a unique and simpler GC profile.

In conclusion, we propose a classification for capturing gene expression plasticity based on the underlying GC profile. Indeed, different GC profiles can drive directly, or by recruiting active mechanisms, concomitant divergences in nucleosome organization. How the relationship between GC profiles and nucleosome organization is established requires further investigations.

### 5.3 MNase-Short-ChIP-seq

Part III asked if MNase digestion level bias can affect the measurement of chromatin-related features. Therefore, MNase-Short-ChIP-seq was developed to include a MNase short digestion as a chromatin shearing approach. This method was then compared to two of the most common shearing approaches: sonication and MNase typical digestion. To test and validate MNase-Short-ChIP-seq, four histone marks were pulled-down: H3K4me3, H3K9ac, H3K18ac and H3K27ac. MNase-Short-ChIP-seq showed higher preservation of MNase-sensitive nucleosomes in the starting material. Therefore, this approach permits to better probe histone PTMs content on these nucleosomes.

High signal-to-noise ratio over active chromatin regions was obtained in MNase-Short-ChIP-seq, also in the starting material. However, the method was confirmed as specific, since histone marks enrichment was not detected closer to inactive genes. As further confirmation, our method captured known enrichments, such as H3K4me3 at the +1 nucleosome, demonstrated in *D.melanogaster* and other species (Howe et al., 2017; Kharchenko et al., 2011), and higher enrichment for H3K9ac at the same position for the NP cluster (Kratz et al., 2010). Interestingly, histone marks pull-downs showed different di-nucleotide profiles. H3K18ac and H3K27ac showed higher WW and lower SS frequencies, indicating higher content of MNase-sensitive nucleosomes, and vice versa for H3K4me3, which is indeed enriched over nucleosome positions with higher GC content. Similar di-nucleotide profile for H3K4me3 has been also described in yeast and human (Tolstorukov et al., 2009).

MNase-Short-ChIP-seq revealed that histone marks deposition around promoters can be influenced by the activity and distance of closer genes. The activity context is discussed in paragraph 5.2.4, since mostly associated with the divergent organization. Regarding the distance context, a subtle shift from higher H3K4me3 to higher acetylation mark occupancies at the -1 position was measured with the increase of gene distances. This result was also found statistically significant at single gene level, in which enrichment for H3K18ac and H3K27ac over H3K4me3 at the -1 position of distant genes was observed (data not shown). Genes closer to each other could not require high acetylation over the intergenic region, since chromatin can be opened by the activity of neighbor genes. In contrast, isolated genes could rely more on the

acetylation marks deposition around their promoters to weaken inter-nucleosome interactions and therefore increase local chromatin accessibility.

In conclusion, we found that MNase-Short-ChIP-seq improves the assessment of chromatin-related features on active chromatin regions. Moreover, thanks to this method, our study reports that deposition of histone tail PTMs around promoters is influenced by the activity and distance of closer genes.

## 6 OUTLOOK

In this study we developed a new method to distinguish nucleosome populations and to characterize fragile and resistant nucleosomes along the genome. However, it could be interesting to study nucleosome fragility and resistance through other experimental approaches, such as chemical mapping, which would require genomic engineering of the H4 gene to introduce the H4S47C mutation. Some attempts were made in mouse embryonic stem cells (Voong et al., 2016), but not all H4 gene copies were modified. This created a mixed pool of wild-type and mutant H4 cells, which could lead to misinterpretations of the data like confusing histone turn-over with nucleosome fragility. To avoid this eventuality, a *Drosophila* strain or cell line with a single copy of H4, containing the required mutation, can be a powerful MNase-independent tool to validate nucleosome populations with an alternative method.

A deeper investigation on what are the activity components that amplify or induce nucleosome fragility around active CREs in *D.melanogaster* is necessary. MNase-Short-ChIP-seq of H3.3, H2A.v and other histone variants on wild type cells and after their knock-downs could be helpful. In a parallel project, we are characterizing nucleosome fragility and resistance around CREs that are responsive to the ecdysone stimulus. For this purpose, we are applying knock-down of the ecdysone receptor and broad, two pivotal factors of the ecdysone cascade. This approach can be extended to a broader spectrum of proteins. Targets can be specific subunits of TFIID, PIC and pausing complexes, or subunits of ATP dependent chromatin remodelers. Moreover, the relationship between pioneer factors and chromatin remodelers with the switching from nucleosome resistance to fragility in enhancers should be better investigated. *D.melanogaster* zygotic genome activation (ZGA) can be an ideal model in this case. Indeed, a strong correlation between chromatin accessibility and Zelda protein (zld) binding events, which most likely acts as pioneer factor, was observed during ZGA (Harrison et al., 2011; Sun et al., 2015). Following the same principles, also the duality between GAF and M1BP factors in the determination of the chromatin organization around promoters can be investigated as well as their relationship with Pol II pausing efficiency. This line of investigation was already pursued in *D.melanogaster*, but without considering nucleosome fragility and resistance (Fuda et al., 2015).

Finally, to better measure the role of Pol II on nucleosome fragility and resistance,  $\alpha$ -amanitin can be used for its specific inhibition. The same approach can be also applied to investigate the Pol II role in the +1 nucleosome positioning on a GC asymmetric context.

The role of the DNA sequence on nucleosome fragility and resistance also needs further investigation. In this direction, the assessment of nucleosome fragility and resistance after reduction or depletion of activity components can provide new insights. For this purpose, chromatin reconstitution by using only *D.melanogaster* genomic DNA and histones can be a powerful tool. Moreover, with this approach, single factors can be also added to measure their specific role in nucleosome fragility and resistance, as similarly performed to study nucleosome organization around promoters in yeast (Krietenstein et al., 2016). The same principle can be translated *in vivo*. In this case, nucleosome fragility and resistance can be measured in embryo before ZGA. Indeed, *D.melanogaster* offers a great opportunity since the zygotic genome shows significant expression only after the 12th cell cycle (Lee et al., 2014). Thus, the ZGA expression dynamics provides a temporal window in which few maternal provided activity components act on the chromatin, and a more direct relationship between DNA sequence and nucleosome fragility and resistance can be assumed. From the technical point of view, this experiment does not require special efforts. Indeed, a differential MNase-seq protocol that uses low amount of embryos was already optimized (200mg of 2-3hr embryo), and further improvements can be obtained to lower the required amount of starting material. The relationship between DNA sequence and nucleosome fragility and resistance can be also studied through a mutational screening on DNA sequences of specific loci. In this regard, CRISPR/Cas9 can be a powerful tool to integrate landing platforms for other recombinases or to directly introduce mutations in the genome. Targets of this approach can be the GC depletion upstream of the TSS and the asymmetric GC context underlying the +1 position in housekeeping genes, and the GC bump on the TSS of developmental genes. Importantly, *in vivo* studies demand controls to avoid data misinterpretations due to coding sequences and DNA motifs manipulations or introduction of homopolymeric DNA sequences. Finally, a plethora of experiments can be designed to include biophysical studies able to directly assess the DNA-histone interaction strength. This setup can then be used to measure how this strength changes according to variation of the DNA sequence, such as GC content and its ratio between nucleosome halves, and 10 bp di-nucleotide periodicities.

Resistant nucleosomes are still poorly characterized. In this study, a relevant role played by them was observed in developmental genes and enhancers. It could be interesting to compare resistant nucleosome mapping before and after ZGA, and to characterize them into heterochromatic domains, such as peri-centromeric, sub-telomeric and facultative heterochromatin domains. Finally, MNase-ChIP-seq with prolonged digestion levels can be used to better probe chromatin-related features associated with resistant nucleosomes.

To better investigate the function and evolutionary conservation of the gene divergent organization, a comparative analysis among fruit fly species can be performed to study the synteny of divergent genes. In case of gene conservation, further investigations are required to better understand the reasons behind. In this regard, it could be interesting to investigate the connection between expression level and nucleosome organization by modifying the distance between TSSs.

Moreover, the divergent gene organization shows several similarities with the antisense transcription that is commonly found in mammalian. The antisense transcription consists of a non-coding RNA that is transcribed from a promoter located upstream of a protein coding gene (Core et al., 2008). In addition, a relationship between antisense and sense TSSs distance with the nucleosome organization was also found, and higher gene activity was measured with the increase of the TSSs distance (Scruggs et al., 2015). In this case, gene activity was assessed by analyzing gene expression level, TF binding events and deposition of the H3K27ac. In our study, most of these features were also reported as typical of the divergent organization. Therefore, an evolutionary relationship between the divergent organization and antisense transcription should be investigated.

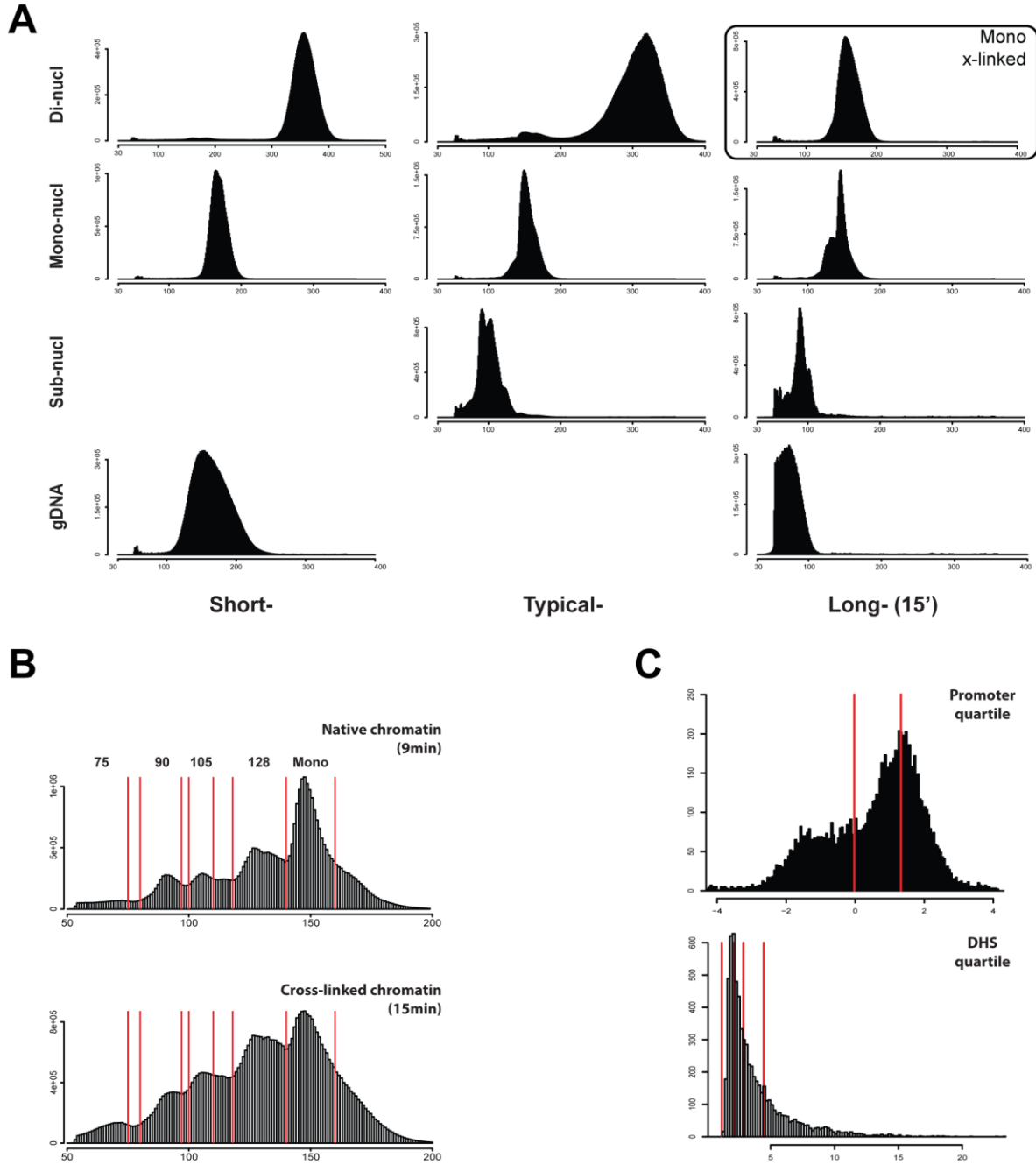
Lastly, the GC content profile captures better the dualism between housekeeping and developmental genes than promoter shape. Therefore, it could be interesting to cluster genes according to the GC content profile to study how this sequence signature is related to gene functionality, nucleosome organization, motif content, chromatin-related features and activity components.

## **7 APPENDIX**

### **7.1 Supplementary Figures**

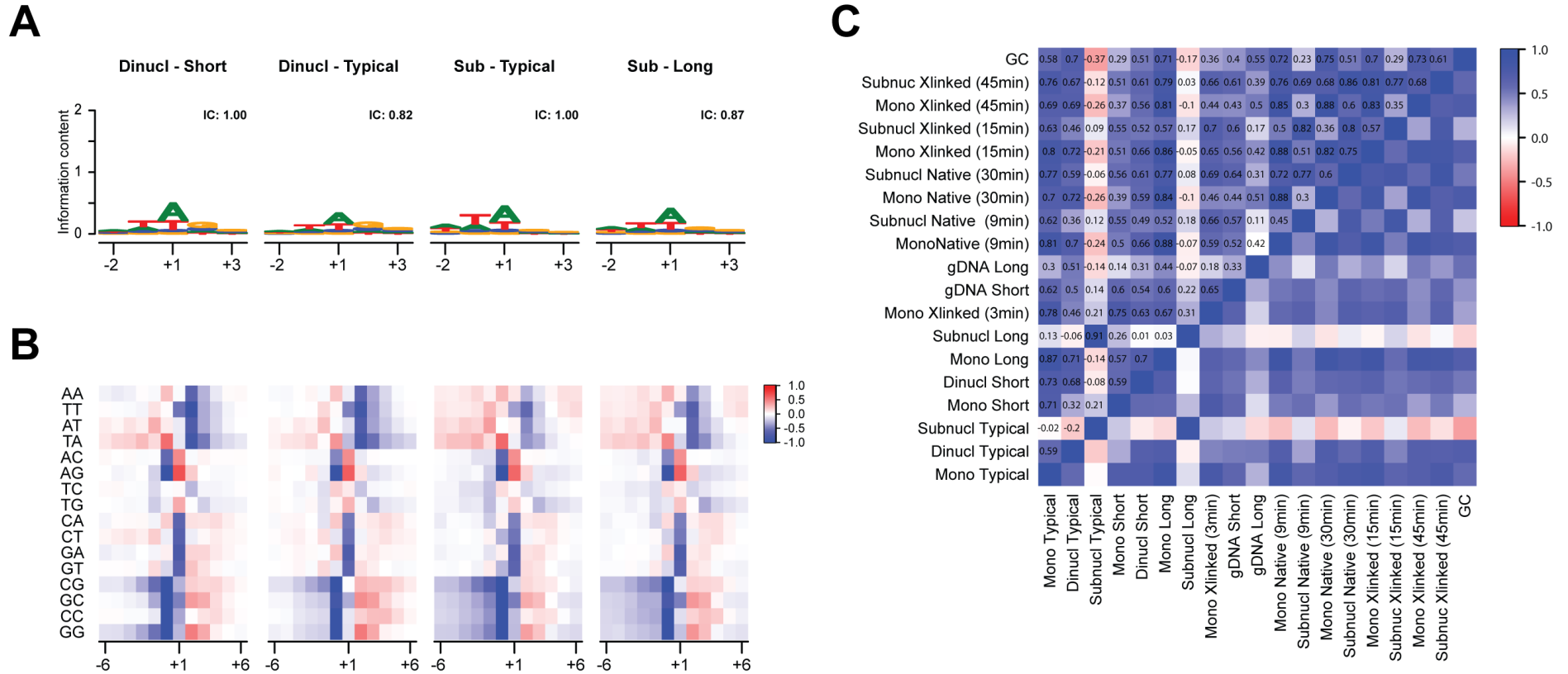


## Figure S1



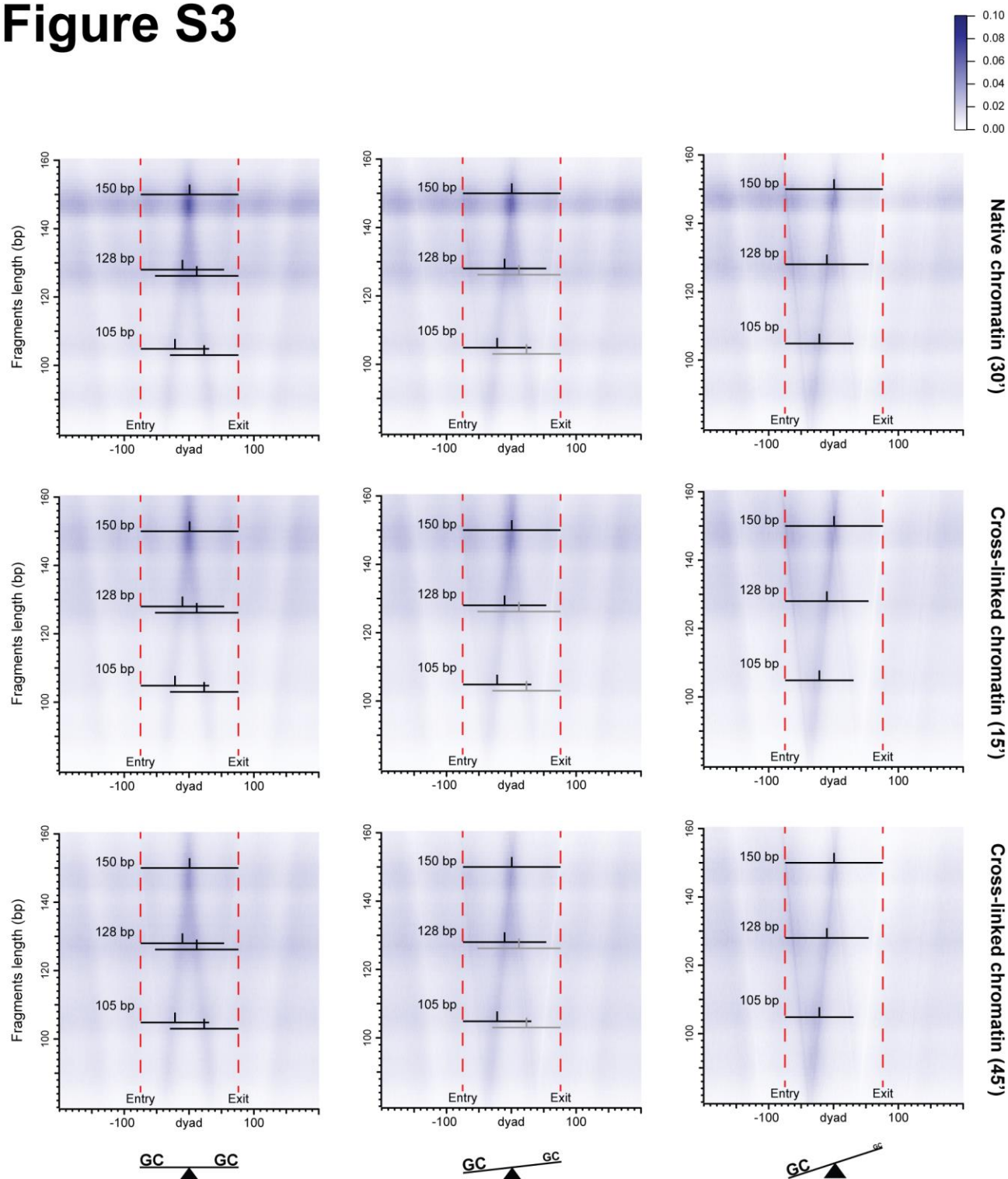
**Figure S1: Fragment length distribution and quartile separation.** (A) Sequenced fragment length distributions at single base pair for all MNase digested chromatin and gDNA. (B) Sequenced fragments distributions for prolonged MNase digestion levels on native and cross-linked chromatin without size-selection between mono- and sub-nucleosome fractions. The red lines indicate the *in silico* size-selection applied to distinguish sub-nucleosome peaks. (C) Gene expression distribution as log<sub>10</sub> (FPKM) (upper panel) and fold change of DHS signal against input within DHS peaks (lower panel). The red lines indicate quartile separation. Quartiles are in decreasing order, so the first contains highest values. In promoter the fourth quartile includes genes in which the RNA-seq data are pretty low or not available.

# Figure S2



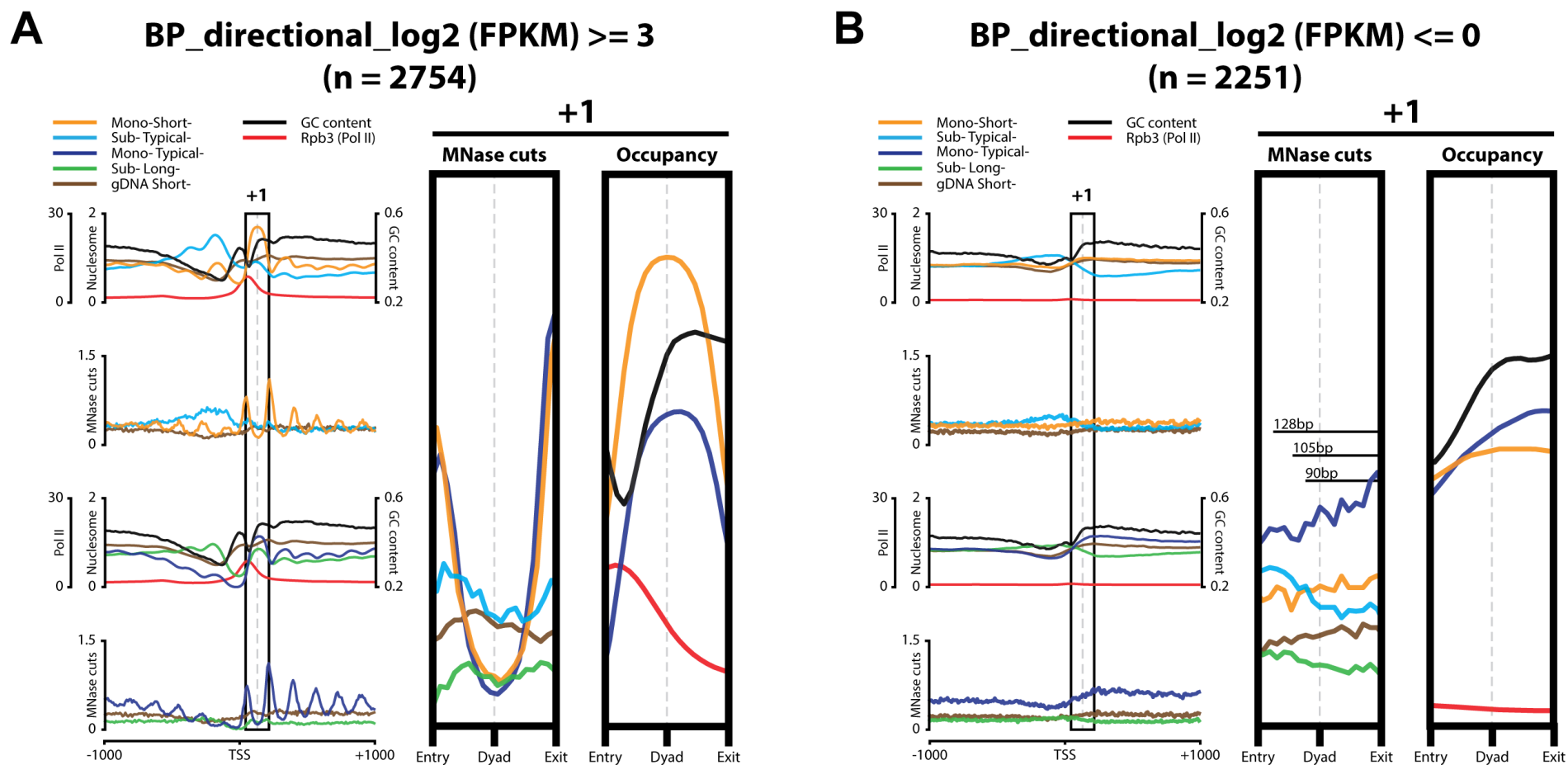
**Figure S2: MNase sequence bias at the cut site (II) and pairwise correlations among samples.** (A) Subfigures matching those in Figure 4.A, but for dinucl- and sub-nucleosome fractions. (B) Subfigures matching those in Figure 4.B, but for the dinucl- and sub-nucleosome fractions. (C) Pairwise Pearson correlations between genome-wide coverage of all sequenced samples. It includes also correlation with the underlying GC content.

## Figure S3



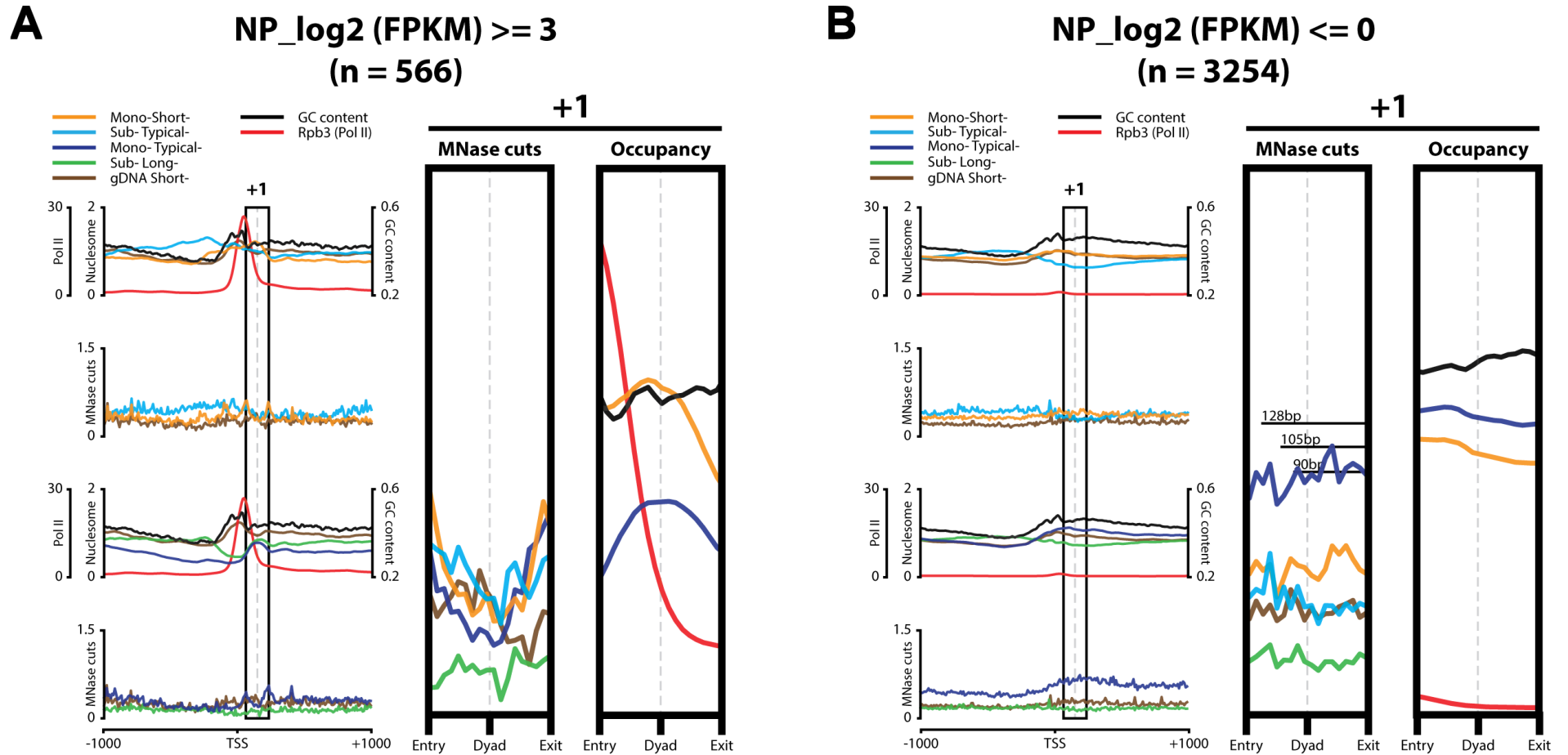
**Figure S3: Asymmetric MNase digestion within the nucleosome (II).** Subfigures matching those in Figure 7.C, but with prolonged digestion level on native chromatin (first line) or from MNase digestion on cross-linked chromatin (second and third lines)

# Figure S4



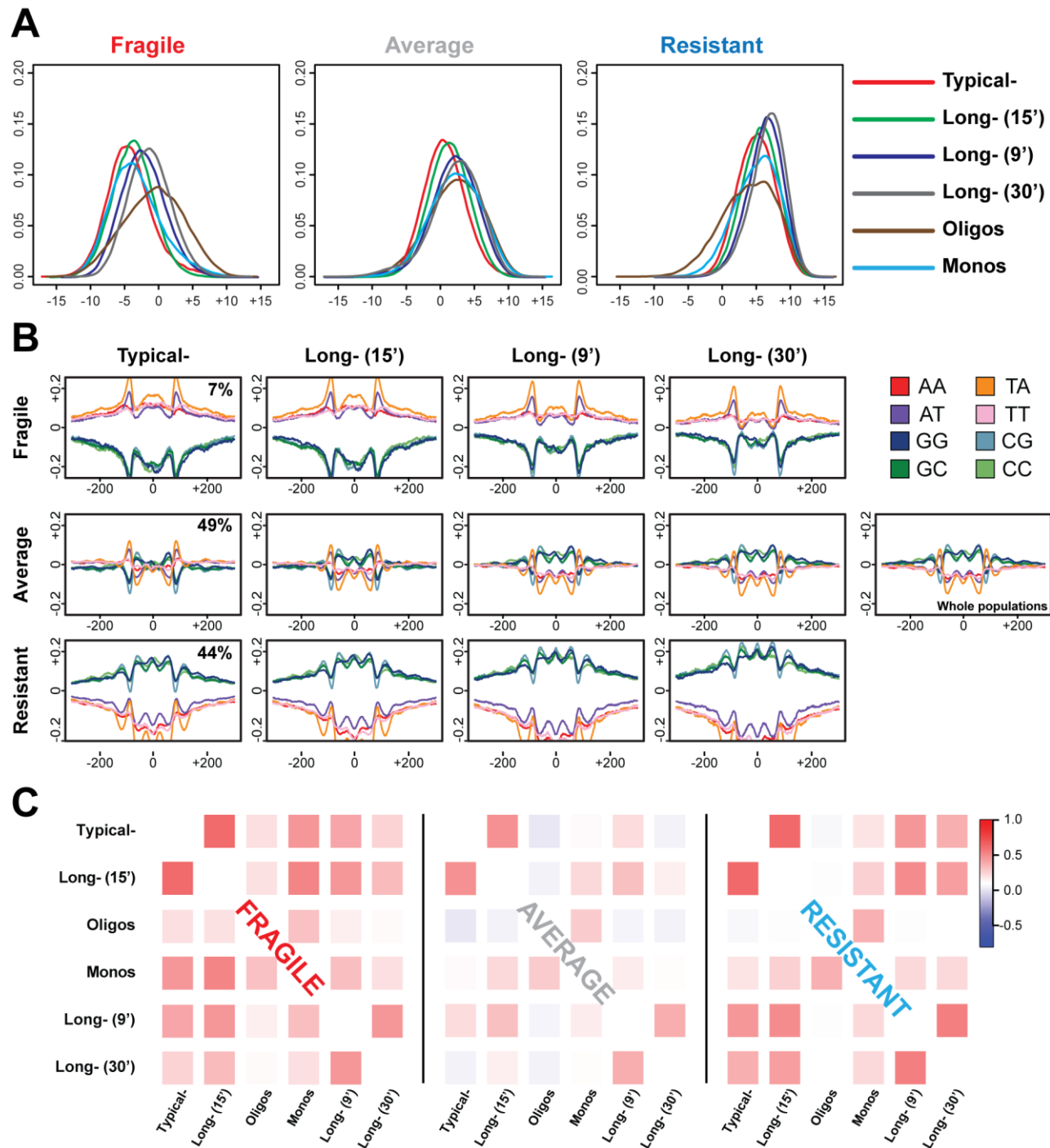
**Figure S4: MNase digestion within the +1 nucleosome (BP directional genes).** (A); (B) Subfigures matching those in Figure 11.A and Figure 11.B, respectively, but considering only genes of the BP directional cluster.

# Figure S5



**Figure S5: MNase digestion within the +1 nucleosome (NP genes).** (A); (B) Subfigures matching those in Figure 11.A and Figure 11.B, respectively, but considering only genes with an NP promoter shape. Identical results were obtained clustering not expressed genes shuffled and reduced in count to match the same n of the expressed gene cluster.

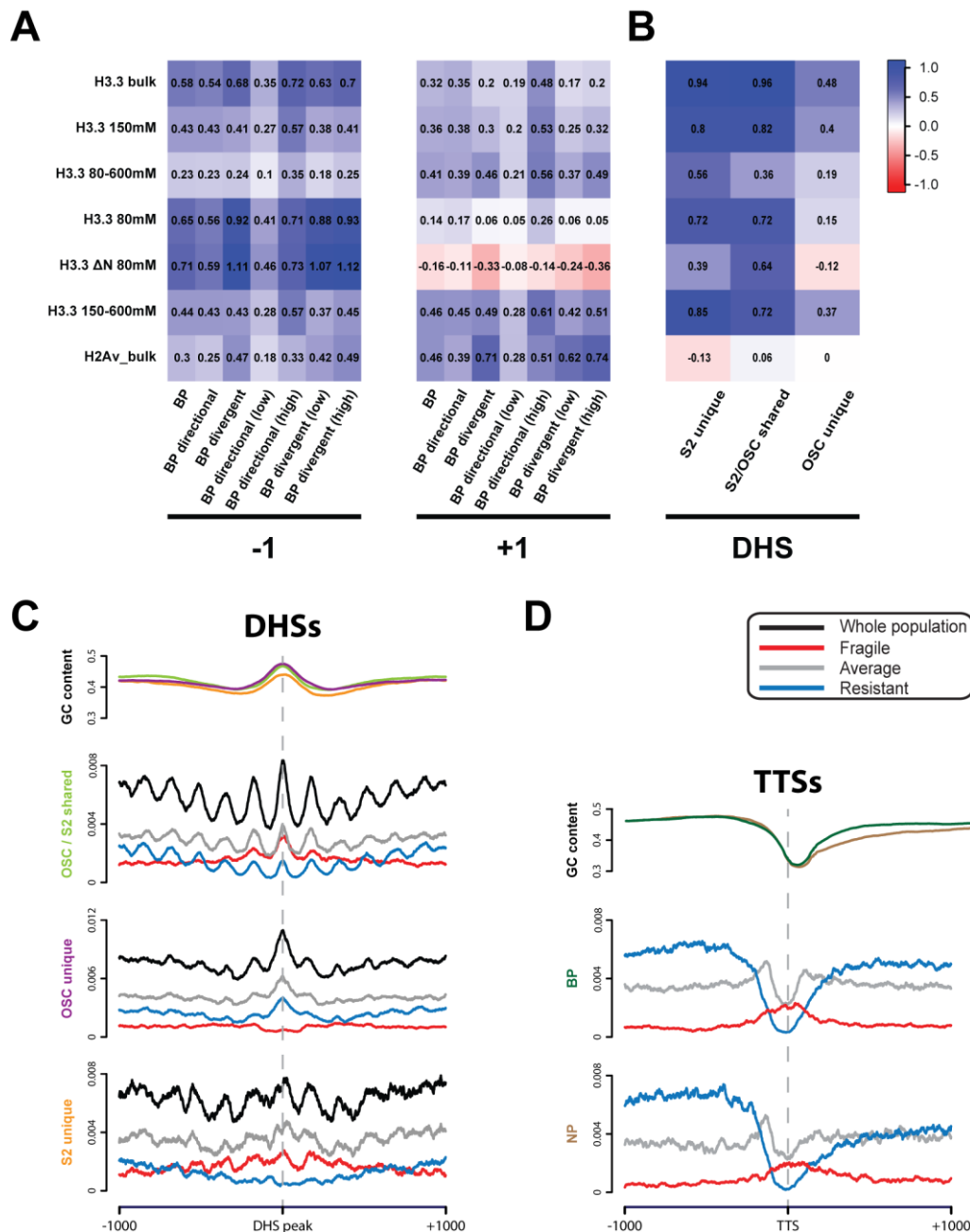
# Figure S6



**Figure S6: Score comparison among nucleosome populations.** (A) Sequence-feature score distributions (see Figure 15.B) among nucleosome populations called with all analyzed scores. (B) Smoothed di-nucleotide enrichment profiles of nucleosome populations aligned by the nucleosome dyad (position 0). On the right of the second line the same it is shown for all called nucleosomes using a nucleR threshold of 0.55. (C) Pairwise Pearson correlations between scores derived from called dyad of fragile (on the left), average (in the middle) and resistant (on the right) nucleosomes.



## Figure S7

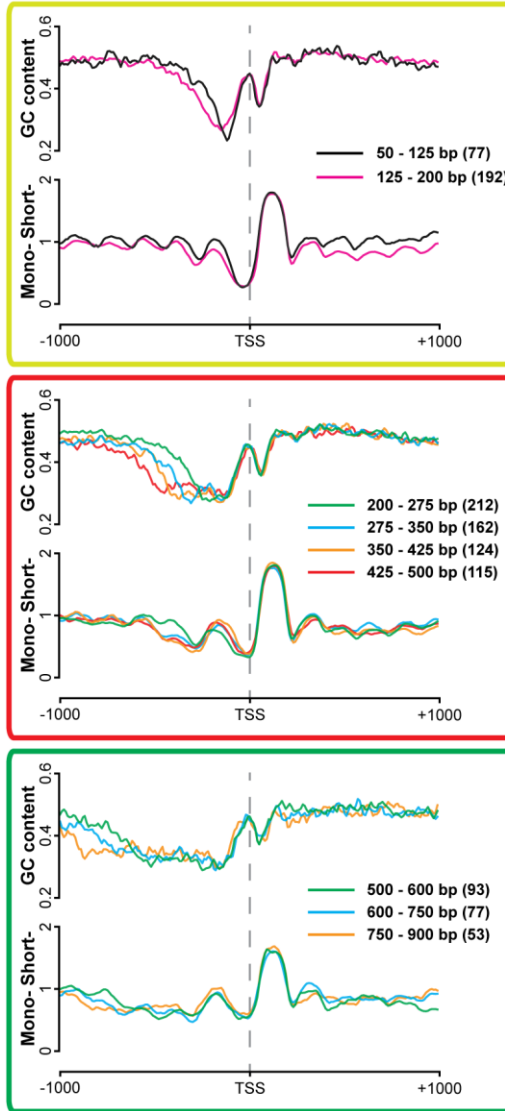


**Figure S7: Histone variant around CREs and nucleosome population around DHS peaks and TTSSs.** (A) Average occupancy of H3.3 and H2A.v derived from salt fractionated chromatin extractions over fixed genomic coordinates corresponded to the likely location of the -1 and +1 nucleosomes around the TSSs. The -1 was positioned at -150 bp upstream of the TSS covering a +/-73 bp region. The +1 was located at +120 bp downstream of the TSS covering +/-73 bp region. Data were collected from this study: (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13217>). (B) Same as (A), but within DHS peaks uniquely open in S2 or OSC cell lines or shared between them. (C) Subfigures matching those in Figure 16, but centered at DHS peaks uniquely open in S2, in OSC cell lines and shared between them. (D) Same as (C), but centered at the TTSSs of BP and NP genes.

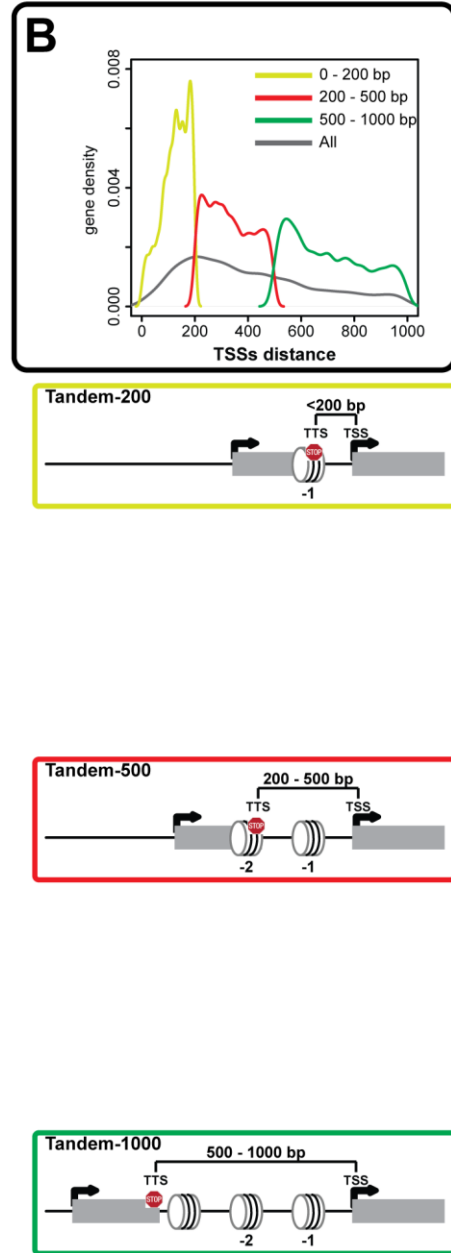
# Figure S8

**A**

**BP tandem -  
First / Second quartiles**

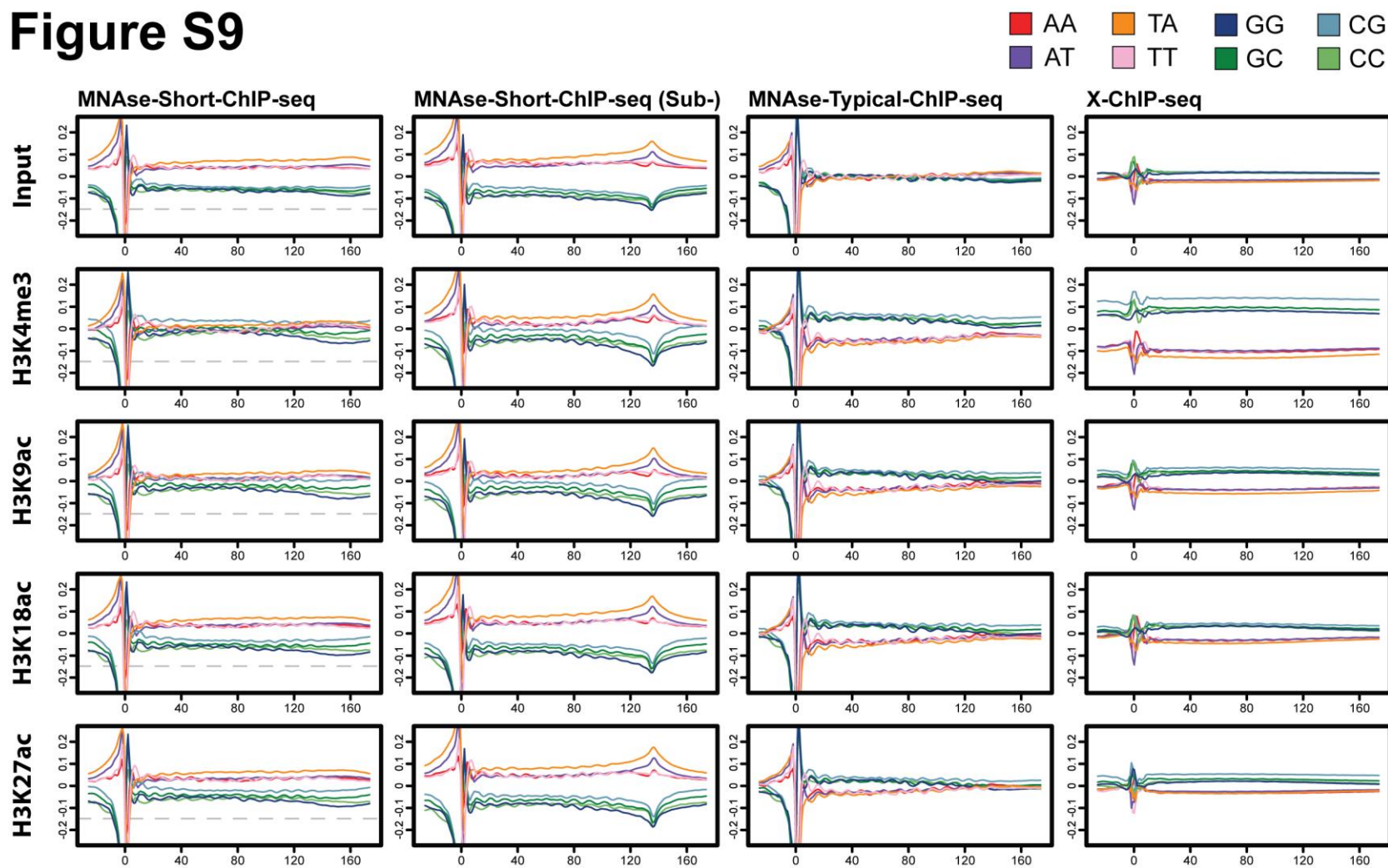


**B**



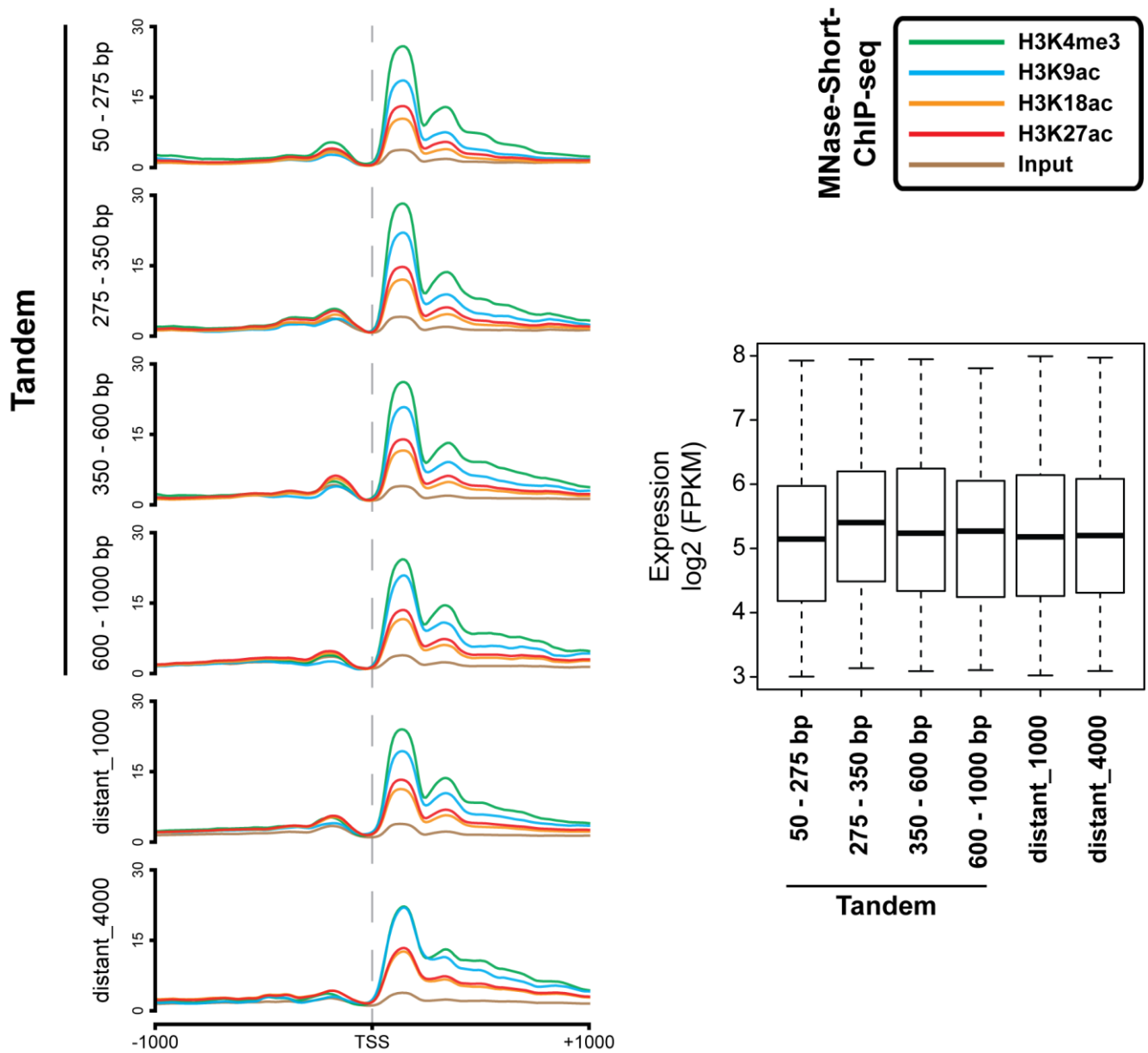
**Figure S8: Relation between TSS/TTS distance with GC content and chromatin profiles in tandem genes.** Subfigure matching Figure 20. (A) Profiles of GC content (25 bp bins) (first line of each panel) and mean coverage of mono- short- (second line of each panel) around BP promoters of tandem genes from the first and second quartiles. The legend in each panel shows the TSS-TTS distance binning and the gene count in brackets. Tandem genes with a TSS-TTS distance between 50 – 200 bp are shown in the upper panel (yellow). Tandem genes with a distance between 200 – 500 bp are shown in the middle panel (red). Tandem genes with a distance between 500 – 1000 bp are shown in the lower panel (green). (B) Smoothed gene density distribution in function of the TSS-TTS distance among tandem genes. The gray line is the distribution for all tandem genes with a TSS-TTS distance between 0 – 1000 bp. Colored distributions match the distance distinction used in (A).



**Figure S9**

**Figure S9: Di-nucleotide profiles of inputs and histone tail PTMs pull-downs from different shearing approaches.** Smoothed di-nucleotide enrichments over the fragment region aligned by the left MNase cut site (see Figure 4) of input (first line) and histone tail PTMs pull-downs (from the second to the fifth lines). Different shearing approaches used in this study are divided in columns: MNase short-digestion (first two columns); MNase typical-digestion (third column); Covaris sonication (fourth column). The second column shows the di-nucleotide enrichment for sub-nucleosomes (55 – 139 bp) derived from the MNase short-digestion approach.

# Figure S10



**Figure S10: Distance context in the histone tail PTMs deposition around the TSS.** On the left, it is shown the mean coverage of input and histone tail PTMs pull-downs from MNase-ChIP-short-seq around the TSS of tandem genes, binned on the TSS-TTS distance, distant\_1000 and distant\_4000 clusters. On the right, it is attached an expression level distribution boxplot (log2 (FPKM)) of each cluster. For a better comparison among them only genes with an expression level of log2 (FPKM)  $\geq 3$  &  $\leq 8$  are considered.

## 7.2 Abbreviations

bp	base-pair
BP	broad-peak promoter
CAGE	cap analysis of gene expression
CGI	CpG island
ChIP	chromatin immunoprecipitation
CpG	cytosine-phosphate-guanine
CRE	<i>cis</i> -regulatory element
DHS	DNase I hypersensitive site
DPE	downstream promoter element
DSIF	DRB sensitivity-inducing factor
EcR	ecdysone receptor
FA	formaldehyde
FPKM	fragments per kilobase million
FRET	Forster resonance energy transfer
GAF	GAGA binding protein
GC content	G + C content
GRF	general regulatory factor
GTF	general transcription factor
Inr	initiation
Kb	kilobase
M1BP	motif 1 binding protein
MNase	micrococcal nuclease
mRNA	messenger RNA
MTE	motif ten elements
N	nucleotide
NDR	nucleosome depleted region
NELF	negative elongation factor
NGS	next-generation sequencing
NP	narrow-peak promoter

## Appendix

o.n.	over night
OSC	ovary stem cells
PIC	pre-initiation complex
Pol II	RNA polymerase II
PTM	post-translational modification
S	G or C nucleotides
S2	<i>Drosophila</i> Schneider 2 cells
SGD	salt gradient dialysis
TAF	TBP associated factors
TBP	TATA-box binding protein
TF	transcription factor
TFBS	transcription factor binding site
TSS	transcription start site
TTS	transcription termination site
UAS	upstream activating sequence
W	A or T nucleotides
ZGA	zygotic genome activation

## 8 BIBLIOGRAPHY

- Adachi, N., and Lieber, M.R. (2002). Bidirectional gene organization: a common architectural feature of the human genome. *Cell* **109**, 807-809.
- Adelman, K., and Lis, J.T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nature reviews Genetics* **13**, 720-731.
- Albert, I., Mavrich, T.N., Tomsho, L.P., Qi, J., Zanton, S.J., Schuster, S.C., and Pugh, B.F. (2007). Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature* **446**, 572-576.
- Aljanabi, S.M., and Martinez, I. (1997). Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic acids research* **25**, 4692-4693.
- Allan, J., Fraser, R.M., Owen-Hughes, T., and Keszenman-Pereyra, D. (2012). Micrococcal nuclease does not substantially bias nucleosome mapping. *Journal of molecular biology* **417**, 152-164.
- Anderson, J.D., and Widom, J. (2000). Sequence and position-dependence of the equilibrium accessibility of nucleosomal DNA target sites. *Journal of molecular biology* **296**, 979-987.
- Arnold, C.D., Gerlach, D., Stelzer, C., Boryn, L.M., Rath, M., and Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074-1077.
- Attrill, H., Falls, K., Goodman, J.L., Millburn, G.H., Antonazzo, G., Rey, A.J., Marygold, S.J., and FlyBase, C. (2016). FlyBase: establishing a Gene Group resource for *Drosophila melanogaster*. *Nucleic acids research* **44**, D786-792.
- Bao, Y., White, C.L., and Luger, K. (2006). Nucleosome core particles containing a poly(dA.dT) sequence element exhibit a locally distorted DNA structure. *Journal of molecular biology* **361**, 617-624.
- Beh, L.Y., Muller, M.M., Muir, T.W., Kaplan, N., and Landweber, L.F. (2015). DNA-guided establishment of nucleosome patterns within coding regions of a eukaryotic genome. *Genome research* **25**, 1727-1738.
- Bintu, L., Ishibashi, T., Dangkulwanich, M., Wu, Y.Y., Lubkowska, L., Kashlev, M., and Bustamante, C. (2012). Nucleosomal elements that control the topography of the barrier to transcription. *Cell* **151**, 738-749.
- Brogaard, K., Xi, L., Wang, J.P., and Widom, J. (2012). A map of nucleosome positions in yeast at base-pair resolution. *Nature* **486**, 496-501.
- Brown, J.B., Boley, N., Eisman, R., May, G.E., Stoiber, M.H., Duff, M.O., Booth, B.W., Wen, J., Park, S., Suzuki, A.M., *et al.* (2014). Diversity and dynamics of the *Drosophila* transcriptome. *Nature* **512**, 393-399.
- Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods* **10**, 1213-1218.
- Cairns, B.R. (2009). The logic of chromatin architecture and remodelling at promoters. *Nature* **461**, 193-198.

- Calo, E., and Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? *Molecular cell* **49**, 825-837.
- Chen, Y., Tokuda, J.M., Topping, T., Meisburger, S.P., Pabit, S.A., Gloss, L.M., and Pollack, L. (2017). Asymmetric unwrapping of nucleosomal DNA propagates asymmetric opening and dissociation of the histone core. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 334-339.
- Chereji, R.V., Kan, T.W., Grudniewska, M.K., Romashchenko, A.V., Berezikov, E., Zhimulev, I.F., Guryev, V., Morozov, A.V., and Moshkin, Y.M. (2016). Genome-wide profiling of nucleosome sensitivity and chromatin accessibility in *Drosophila melanogaster*. *Nucleic acids research* **44**, 1036-1051.
- Chereji, R.V., Ocampo, J., and Clark, D.J. (2017). MNase-Sensitive Complexes in Yeast: Nucleosomes and Non-histone Barriers. *Molecular cell* **65**, 565-577 e563.
- Chua, E.Y., Vasudevan, D., Davey, G.E., Wu, B., and Davey, C.A. (2012). The mechanics behind DNA sequence-dependent properties of the nucleosome. *Nucleic acids research* **40**, 6338-6352.
- Chung, H.R., Dunkel, I., Heise, F., Linke, C., Krobisch, S., Ehrenhofer-Murray, A.E., Sperling, S.R., and Vingron, M. (2010). The effect of micrococcal nuclease digestion on nucleosome positioning data. *PLoS one* **5**, e15754.
- Cockell, M., Rhodes, D., and Klug, A. (1983). Location of the primary sites of micrococcal nuclease cleavage on the nucleosome core. *Journal of molecular biology* **170**, 423-446.
- Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845-1848.
- Cui, F., Cole, H.A., Clark, D.J., and Zhurkin, V.B. (2012). Transcriptional activation of yeast genes disrupts intragenic nucleosome phasing. *Nucleic acids research* **40**, 10753-10764.
- Davey, C.A., Sargent, D.F., Luger, K., Maeder, A.W., and Richmond, T.J. (2002). Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *Journal of molecular biology* **319**, 1097-1113.
- Deniz, O., Flores, O., Battistini, F., Perez, A., Soler-Lopez, M., and Orozco, M. (2011). Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast. *BMC genomics* **12**, 489.
- Dingwall, C., Lomonosoff, G.P., and Laskey, R.A. (1981). High sequence specificity of micrococcal nuclease. *Nucleic acids research* **9**, 2659-2673.
- Drew, H.R., and Travers, A.A. (1985). DNA bending and its relation to nucleosome positioning. *Journal of molecular biology* **186**, 773-790.
- Ellis, L.L., Huang, W., Quinn, A.M., Ahuja, A., Alfrejd, B., Gomez, F.E., Hjelman, C.E., Moore, K.L., Mackay, T.F., Johnston, J.S., *et al.* (2014). Intrapopulation genome size variation in *D. melanogaster* reflects life history variation and plasticity. *PLoS genetics* **10**, e1004522.
- Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shores, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., *et al.* (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-49.
- Eslami-Mossallam, B., Schiessel, H., and van Noort, J. (2016). Nucleosome dynamics: Sequence matters. *Advances in colloid and interface science* **232**, 101-113.

- Fan, X., Moqtaderi, Z., Jin, Y., Zhang, Y., Liu, X.S., and Struhl, K. (2010). Nucleosome depletion at yeast terminators is not intrinsic and can occur by a transcriptional mechanism linked to 3'-end formation. *Proceedings of the National Academy of Sciences of the United States of America* 107, 17945-17950.
- Fenouil, R., Cauchy, P., Koch, F., Descostes, N., Cabeza, J.Z., Innocenti, C., Ferrier, P., Spicuglia, S., Gut, M., Gut, I., *et al.* (2012). CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. *Genome research* 22, 2399-2408.
- Field, Y., Kaplan, N., Fondufe-Mittendorf, Y., Moore, I.K., Sharon, E., Lubling, Y., Widom, J., and Segal, E. (2008). Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS computational biology* 4, e1000216.
- Floer, M., Wang, X., Prabhu, V., Berrozpe, G., Narayan, S., Spagna, D., Alvarez, D., Kendall, J., Krasnitz, A., Stepansky, A., *et al.* (2010). A RSC/nucleosome complex determines chromatin architecture and facilitates activator binding. *Cell* 141, 407-418.
- Flores, O., Deniz, O., Soler-Lopez, M., and Orozco, M. (2014). Fuzziness and noise in nucleosomal architecture. *Nucleic acids research* 42, 4934-4946.
- Flores, O., and Orozco, M. (2011). nucleR: a package for non-parametric nucleosome positioning. *Bioinformatics* 27, 2149-2150.
- Fu, Y., Sinha, M., Peterson, C.L., and Weng, Z. (2008). The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS genetics* 4, e1000138.
- Fuda, N.J., Guertin, M.J., Sharma, S., Danko, C.G., Martins, A.L., Siepel, A., and Lis, J.T. (2015). GAGA factor maintains nucleosome-free regions and has a role in RNA polymerase II recruitment to promoters. *PLoS genetics* 11, e1005108.
- Fuda, N.J., and Lis, J.T. (2013). A new player in Pol II pausing. *The EMBO journal* 32, 1796-1798.
- Gaffney, D.J., McVicker, G., Pai, A.A., Fondufe-Mittendorf, Y.N., Lewellen, N., Michelini, K., Widom, J., Gilad, Y., and Pritchard, J.K. (2012). Controls of nucleosome positioning in the human genome. *PLoS genetics* 8, e1003036.
- Gordon, F., Luger, K., and Hansen, J.C. (2005). The core histone N-terminal tail domains function independently and additively during salt-dependent oligomerization of nucleosomal arrays. *The Journal of biological chemistry* 280, 33701-33706.
- Gossett, A.J., and Lieb, J.D. (2012). In vivo effects of histone H3 depletion on nucleosome occupancy and position in *Saccharomyces cerevisiae*. *PLoS genetics* 8, e1002771.
- Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri, C.G., van Baren, M.J., Boley, N., Booth, B.W., *et al.* (2011). The developmental transcriptome of *Drosophila melanogaster*. *Nature* 471, 473-479.
- Haberle, V., and Lenhard, B. (2016). Promoter architectures and developmental gene regulation. *Seminars in cell & developmental biology* 57, 11-23.
- Hall, M.A., Shundrovsky, A., Bai, L., Fulbright, R.M., Lis, J.T., and Wang, M.D. (2009). High-resolution dynamic mapping of histone-DNA interactions in a nucleosome. *Nature structural & molecular biology* 16, 124-129.
- Harrison, M.M., Li, X.Y., Kaplan, T., Botchan, M.R., and Eisen, M.B. (2011). Zelda binding in the early *Drosophila melanogaster* embryo marks regions subsequently activated at the maternal-to-zygotic transition. *PLoS genetics* 7, e1002266.

- Hasson, D., Panchenko, T., Salimian, K.J., Salman, M.U., Sekulic, N., Alonso, A., Warburton, P.E., and Black, B.E. (2013). The octamer is the major form of CENP-A nucleosomes at human centromeres. *Nature structural & molecular biology* 20, 687-695.
- Henikoff, J.G., Belsky, J.A., Krassovsky, K., MacAlpine, D.M., and Henikoff, S. (2011). Epigenome characterization at single base-pair resolution. *Proceedings of the National Academy of Sciences of the United States of America* 108, 18318-18323.
- Henikoff, S., Henikoff, J.G., Sakai, A., Loeb, G.B., and Ahmad, K. (2009). Genome-wide profiling of salt fractions maps physical properties of chromatin. *Genome research* 19, 460-469.
- Ho, J.W., Jung, Y.L., Liu, T., Alver, B.H., Lee, S., Ikegami, K., Sohn, K.A., Minoda, A., Tolstorukov, M.Y., Appert, A., *et al.* (2014). Comparative analysis of metazoan chromatin organization. *Nature* 512, 449-452.
- Horz, W., and Altenburger, W. (1981). Sequence specific cleavage of DNA by micrococcal nuclease. *Nucleic acids research* 9, 2643-2658.
- Hoskins, R.A., Landolin, J.M., Brown, J.B., Sandler, J.E., Takahashi, H., Lassmann, T., Yu, C., Booth, B.W., Zhang, D., Wan, K.H., *et al.* (2011). Genome-wide analysis of promoter architecture in *Drosophila melanogaster*. *Genome research* 21, 182-192.
- Howe, F.S., Fischl, H., Murray, S.C., and Mellor, J. (2017). Is H3K4me3 instructive for transcription activation? *BioEssays : news and reviews in molecular, cellular and developmental biology* 39, 1-12.
- Ishii, H., Kadonaga, J.T., and Ren, B. (2015). MPE-seq, a new method for the genome-wide analysis of chromatin structure. *Proceedings of the National Academy of Sciences of the United States of America* 112, E3457-3465.
- Iwafuchi-Doi, M., Donahue, G., Kakumanu, A., Watts, J.A., Mahony, S., Pugh, B.F., Lee, D., Kaestner, K.H., and Zaret, K.S. (2016). The Pioneer Transcription Factor FoxA Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation. *Molecular cell* 62, 79-91.
- Iwasaki, W., Miya, Y., Horikoshi, N., Osakabe, A., Taguchi, H., Tachiwana, H., Shibata, T., Kagawa, W., and Kurumizaka, H. (2013). Contribution of histone N-terminal tails to the structure and stability of nucleosomes. *FEBS open bio* 3, 363-369.
- Jeffers, T.E., and Lieb, J.D. (2017). Nucleosome fragility is associated with future transcriptional response to developmental cues and stress in *C. elegans*. *Genome research* 27, 75-86.
- Jin, C., and Felsenfeld, G. (2007). Nucleosome stability mediated by histone variants H3.3 and H2A.Z. *Genes & development* 21, 1519-1529.
- Jin, C., Zang, C., Wei, G., Cui, K., Peng, W., Zhao, K., and Felsenfeld, G. (2009). H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. *Nature genetics* 41, 941-945.
- Juven-Gershon, T., and Kadonaga, J.T. (2010). Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Developmental biology* 339, 225-229.
- Kaplan, N., Hughes, T.R., Lieb, J.D., Widom, J., and Segal, E. (2010a). Contribution of histone sequence preferences to nucleosome organization: proposed definitions and methodology. *Genome biology* 11, 140.



- Kaplan, N., Moore, I., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., Hughes, T.R., Lieb, J.D., Widom, J., and Segal, E. (2010b). Nucleosome sequence preferences influence in vivo nucleosome organization. *Nature structural & molecular biology* *17*, 918-920.
- Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., LeProust, E.M., Hughes, T.R., Lieb, J.D., Widom, J., *et al.* (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* *458*, 362-366.
- Kasinathan, S., Orsi, G.A., Zentner, G.E., Ahmad, K., and Henikoff, S. (2014). High-resolution mapping of transcription factor binding sites on native chromatin. *Nature methods* *11*, 203-209.
- Kharchenko, P.V., Alekseyenko, A.A., Schwartz, Y.B., Minoda, A., Riddle, N.C., Ernst, J., Sabo, P.J., Larschan, E., Gorchakov, A.A., Gu, T., *et al.* (2011). Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature* *471*, 480-485.
- Khuong, M.T., Fei, J., Ishii, H., and Kadonaga, J.T. (2015). Prenucleosomes and Active Chromatin. *Cold Spring Harbor symposia on quantitative biology* *80*, 65-72.
- Koopmans, W.J., Buning, R., Schmidt, T., and van Noort, J. (2009). spFRET using alternating excitation and FCS reveals progressive DNA unwrapping in nucleosomes. *Biophysical journal* *97*, 195-204.
- Koopmans, W.J., Schmidt, T., and van Noort, J. (2008). Nucleosome immobilization strategies for single-pair FRET microscopy. *Chemphyschem : a European journal of chemical physics and physical chemistry* *9*, 2002-2009.
- Kornberg, R.D., and Stryer, L. (1988). Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. *Nucleic acids research* *16*, 6677-6690.
- Kratz, A., Arner, E., Saito, R., Kubosaki, A., Kawai, J., Suzuki, H., Carninci, P., Arakawa, T., Tomita, M., Hayashizaki, Y., *et al.* (2010). Core promoter structure and genomic context reflect histone 3 lysine 9 acetylation patterns. *BMC genomics* *11*, 257.
- Krietenstein, N., Wal, M., Watanabe, S., Park, B., Peterson, C.L., Pugh, B.F., and Korber, P. (2016). Genomic Nucleosome Organization Reconstituted with Pure Proteins. *Cell* *167*, 709-721 e712.
- Kubik, S., Bruzzone, M.J., Albert, B., and Shore, D. (2017a). A Reply to "MNase-Sensitive Complexes in Yeast: Nucleosomes and Non-histone Barriers," by Chereji *et al.* *Molecular cell* *65*, 578-580.
- Kubik, S., Bruzzone, M.J., Jacquet, P., Falcone, J.L., Rougemont, J., and Shore, D. (2015). Nucleosome Stability Distinguishes Two Different Promoter Types at All Protein-Coding Genes in Yeast. *Molecular cell* *60*, 422-434.
- Kubik, S., Bruzzone, M.J., and Shore, D. (2017b). Establishing nucleosome architecture and stability at promoters: Roles of pioneer transcription factors and the RSC chromatin remodeler. *BioEssays : news and reviews in molecular, cellular and developmental biology* *39*.
- Kundaje, A., Kyriazopoulou-Panagiotopoulou, S., Libbrecht, M., Smith, C.L., Raha, D., Winters, E.E., Johnson, S.M., Snyder, M., Batzoglou, S., and Sidow, A. (2012). Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome research* *22*, 1735-1747.
- Kwak, H., Fuda, N.J., Core, L.J., and Lis, J.T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* *339*, 950-953.
- Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods* *9*, 357-359.

- Lantermann, A.B., Straub, T., Stralfors, A., Yuan, G.C., Ekwall, K., and Korber, P. (2010). *Schizosaccharomyces pombe* genome-wide nucleosome mapping reveals positioning mechanisms distinct from those of *Saccharomyces cerevisiae*. *Nature structural & molecular biology* 17, 251-257.
- Lee, M.T., Bonneau, A.R., and Giraldez, A.J. (2014). Zygotic genome activation during the maternal-to-zygotic transition. *Annual review of cell and developmental biology* 30, 581-613.
- Lee, W., Tillo, D., Bray, N., Morse, R.H., Davis, R.W., Hughes, T.R., and Nislow, C. (2007). A high-resolution atlas of nucleosome occupancy in yeast. *Nature genetics* 39, 1235-1244.
- Li, G., Levitus, M., Bustamante, C., and Widom, J. (2005). Rapid spontaneous accessibility of nucleosomal DNA. *Nature structural & molecular biology* 12, 46-53.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078-2079.
- Li, J., and Gilmour, D.S. (2013). Distinct mechanisms of transcriptional pausing orchestrated by GAGA factor and M1BP, a novel transcription factor. *The EMBO journal* 32, 1829-1841.
- Li, X.Q., and Du, D. (2014). Variation, evolution, and correlation analysis of C+G content and genome or chromosome size in different kingdoms and phyla. *PloS one* 9, e88339.
- Lin, J.M., Collins, P.J., Trinklein, N.D., Fu, Y., Xi, H., Myers, R.M., and Weng, Z. (2007). Transcription factor binding and modified histones in human bidirectional promoters. *Genome research* 17, 818-827.
- Liu, H., Zhang, R., Xiong, W., Guan, J., Zhuang, Z., and Zhou, S. (2014). A comparative evaluation on prediction methods of nucleosome positioning. *Briefings in bioinformatics* 15, 1014-1027.
- Liu, X., Kraus, W.L., and Bai, X. (2015). Ready, pause, go: regulation of RNA polymerase II pausing and release by cellular signaling pathways. *Trends in biochemical sciences* 40, 516-525.
- Locke, G., Tolkunov, D., Moqtaderi, Z., Struhl, K., and Morozov, A.V. (2010). High-throughput sequencing reveals a simple model of nucleosome energetics. *Proceedings of the National Academy of Sciences of the United States of America* 107, 20998-21003.
- Lowary, P.T., and Widom, J. (1998). New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *Journal of molecular biology* 276, 19-42.
- Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* 389, 251-260.
- Mapendano, C.K., Lykke-Andersen, S., Kjems, J., Bertrand, E., and Jensen, T.H. (2010). Crosstalk between mRNA 3' end processing and transcription initiation. *Molecular cell* 40, 410-422.
- Mavrich, T.N., Ioshikhes, I.P., Venters, B.J., Jiang, C., Tomsho, L.P., Qi, J., Schuster, S.C., Albert, I., and Pugh, B.F. (2008a). A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome research* 18, 1073-1083.
- Mavrich, T.N., Jiang, C., Ioshikhes, I.P., Li, X., Venters, B.J., Zanton, S.J., Tomsho, L.P., Qi, J., Glaser, R.L., Schuster, S.C., *et al.* (2008b). Nucleosome organization in the *Drosophila* genome. *Nature* 453, 358-362.
- Meng, H., Andresen, K., and van Noort, J. (2015). Quantitative analysis of single-molecule force spectroscopy on folded chromatin fibers. *Nucleic acids research* 43, 3578-3590.

- Mieczkowski, J., Cook, A., Bowman, S.K., Mueller, B., Alver, B.H., Kundu, S., Deaton, A.M., Urban, J.A., Larschan, E., Park, P.J., *et al.* (2016). MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. *Nature communications* 7, 11485.
- Moshkin, Y.M., Chalkley, G.E., Kan, T.W., Reddy, B.A., Ozgur, Z., van Ijcken, W.F., Dekkers, D.H., Demmers, J.A., Travers, A.A., and Verrijzer, C.P. (2012). Remodelers organize cellular chromatin by counteracting intrinsic histone-DNA sequence preferences in a class-specific manner. *Molecular and cellular biology* 32, 675-688.
- Moyle-Heyrman, G., Zaichuk, T., Xi, L., Zhang, Q., Uhlenbeck, O.C., Holmgren, R., Widom, J., and Wang, J.P. (2013). Chemical map of *Schizosaccharomyces pombe* reveals species-specific features in nucleosome positioning. *Proceedings of the National Academy of Sciences of the United States of America* 110, 20158-20163.
- Mueller, B., Mieczkowski, J., Kundu, S., Wang, P., Sadreyev, R., Tolstorukov, M.Y., and Kingston, R.E. (2017). Widespread changes in nucleosome accessibility without changes in nucleosome occupancy during a rapid transcriptional induction. *Genes & development* 31, 451-462.
- Ngo, T.T., Zhang, Q., Zhou, R., Yodh, J.G., and Ha, T. (2015). Asymmetric unwrapping of nucleosomes under tension directed by DNA local flexibility. *Cell* 160, 1135-1144.
- Nikitina, T., Wang, D., Gomberg, M., Grigoryev, S.A., and Zhurkin, V.B. (2013). Combined micrococcal nuclease and exonuclease III digestion reveals precise positions of the nucleosome core/linker junctions: implications for high-resolution nucleosome mapping. *Journal of molecular biology* 425, 1946-1960.
- Polach, K.J., Lowary, P.T., and Widom, J. (2000). Effects of core histone tail domains on the equilibrium constants for dynamic DNA site accessibility in nucleosomes. *Journal of molecular biology* 298, 211-223.
- Polach, K.J., and Widom, J. (1995). Mechanism of protein access to specific DNA sequences in chromatin: a dynamic equilibrium model for gene regulation. *Journal of molecular biology* 254, 130-149.
- Polach, K.J., and Widom, J. (1996). A model for the cooperative binding of eukaryotic regulatory proteins to nucleosomal target sites. *Journal of molecular biology* 258, 800-812.
- Prendergast, J.G., and Semple, C.A. (2011). Widespread signatures of recent selection linked to nucleosome positioning in the human lineage. *Genome research* 21, 1777-1787.
- Proudfoot, N.J. (2016). Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science* 352, aad9926.
- Rach, E.A., Winter, D.R., Benjamin, A.M., Corcoran, D.L., Ni, T., Zhu, J., and Ohler, U. (2011). Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. *PLoS genetics* 7, e1001274.
- Rach, E.A., Yuan, H.Y., Majoros, W.H., Tomancak, P., and Ohler, U. (2009). Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the *Drosophila* genome. *Genome biology* 10, R73.
- Ramachandran, S., Zentner, G.E., and Henikoff, S. (2015). Asymmetric nucleosomes flank promoters in the budding yeast genome. *Genome research* 25, 381-390.
- Ramirez, F., Dundar, F., Diehl, S., Gruning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic acids research* 42, W187-191.

- Raveh-Sadka, T., Levo, M., Shabi, U., Shany, B., Keren, L., Lotan-Pompan, M., Zeevi, D., Sharon, E., Weinberger, A., and Segal, E. (2012). Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nature genetics* 44, 743-750.
- Rhee, H.S., Bataille, A.R., Zhang, L., and Pugh, B.F. (2014). Subnucleosomal structures and nucleosome asymmetry across a genome. *Cell* 159, 1377-1388.
- Richmond, T.J., and Davey, C.A. (2003). The structure of DNA in the nucleosome core. *Nature* 423, 145-150.
- Rizzo, J.M., Bard, J.E., and Buck, M.J. (2012). Standardized collection of MNase-seq experiments enables unbiased dataset comparisons. *BMC molecular biology* 13, 15.
- Robinson, P.J., and Rhodes, D. (2006). Structure of the '30 nm' chromatin fibre: a key role for the linker histone. *Current opinion in structural biology* 16, 336-343.
- Sainsbury, S., Bernecky, C., and Cramer, P. (2015). Structural basis of transcription initiation by RNA polymerase II. *Nature reviews Molecular cell biology* 16, 129-143.
- Satchwell, S.C., Drew, H.R., and Travers, A.A. (1986). Sequence periodicities in chicken nucleosome core DNA. *Journal of molecular biology* 191, 659-675.
- Schor, I.E., Degner, J.F., Harnett, D., Cannavo, E., Casale, F.P., Shim, H., Garfield, D.A., Birney, E., Stephens, M., Stegle, O., *et al.* (2017). Promoter shape varies across populations and affects promoter evolution and expression noise. *Nature genetics* 49, 550-558.
- Scruggs, B.S., Gilchrist, D.A., Nechaev, S., Muse, G.W., Burkholder, A., Fargo, D.C., and Adelman, K. (2015). Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. *Molecular cell* 58, 1101-1112.
- Segal, E., and Widom, J. (2009). Poly(dA:dT) tracts: major determinants of nucleosome organization. *Current opinion in structural biology* 19, 65-71.
- Sikorski, T.W., and Buratowski, S. (2009). The basal initiation machinery: beyond the general transcription factors. *Current opinion in cell biology* 21, 344-351.
- Skene, P.J., Hernandez, A.E., Groudine, M., and Henikoff, S. (2014). The nucleosomal barrier to promoter escape by RNA polymerase II is overcome by the chromatin remodeler Chd1. *eLife* 3, e02042.
- Skourti-Stathaki, K., and Proudfoot, N.J. (2014). A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression. *Genes & development* 28, 1384-1396.
- Struhl, K., and Segal, E. (2013). Determinants of nucleosome positioning. *Nature structural & molecular biology* 20, 267-273.
- Sun, Y., Nien, C.Y., Chen, K., Liu, H.Y., Johnston, J., Zeitlinger, J., and Rushlow, C. (2015). Zelda overcomes the high intrinsic nucleosome barrier at enhancers during *Drosophila* zygotic genome activation. *Genome research* 25, 1703-1714.
- Suter, B., Schnappauf, G., and Thoma, F. (2000). Poly(dA:dT) sequences exist as rigid DNA structures in nucleosome-free yeast promoters in vivo. *Nucleic acids research* 28, 4083-4089.
- Swinstead, E.E., Paakinaho, V., Presman, D.M., and Hager, G.L. (2016). Pioneer factors and ATP-dependent chromatin remodeling factors interact dynamically: A new perspective: Multiple transcription factors can effect chromatin pioneer functions through dynamic interactions with ATP-dependent chromatin remodeling factors. *BioEssays : news and reviews in molecular, cellular and developmental biology* 38, 1150-1157.

- Teif, V.B. (2016). Nucleosome positioning: resources and tools online. *Briefings in bioinformatics* 17, 745-757.
- Teves, S.S., and Henikoff, S. (2011). Heat shock reduces stalled RNA polymerase II and nucleosome turnover genome-wide. *Genes & development* 25, 2387-2397.
- Teves, S.S., and Henikoff, S. (2014). Transcription-generated torsional stress destabilizes nucleosomes. *Nature structural & molecular biology* 21, 88-94.
- Teves, S.S., Weber, C.M., and Henikoff, S. (2014a). Transcribing through the nucleosome. *Trends in biochemical sciences* 39, 577-586.
- Teves, S.S., Weber, C.M., and Henikoff, S. (2014b). Transcribing through the nucleosome. *Trends in biochemical sciences* 39, 577-586.
- Thastrom, A., Lowary, P.T., Widlund, H.R., Cao, H., Kubista, M., and Widom, J. (1999). Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. *Journal of molecular biology* 288, 213-229.
- Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., *et al.* (2012). The accessible chromatin landscape of the human genome. *Nature* 489, 75-82.
- Tillo, D., and Hughes, T.R. (2009). G+C content dominates intrinsic nucleosome occupancy. *BMC bioinformatics* 10, 442.
- Tillo, D., Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Field, Y., Lieb, J.D., Widom, J., Segal, E., and Hughes, T.R. (2010). High nucleosome occupancy is encoded at human regulatory sequences. *PloS one* 5, e9129.
- Tirosh, I., and Barkai, N. (2008). Two strategies for gene regulation by promoter nucleosomes. *Genome research* 18, 1084-1091.
- Tolstorukov, M.Y., Kharchenko, P.V., Goldman, J.A., Kingston, R.E., and Park, P.J. (2009). Comparative analysis of H2A.Z nucleosome organization in the human and yeast genomes. *Genome research* 19, 967-977.
- Trinklein, N.D., Aldred, S.F., Hartman, S.J., Schroeder, D.I., Otilar, R.P., and Myers, R.M. (2004). An abundance of bidirectional promoters in the human genome. *Genome research* 14, 62-66.
- Tsankov, A., Yanagisawa, Y., Rhind, N., Regev, A., and Rando, O.J. (2011). Evolutionary divergence of intrinsic and trans-regulated nucleosome positioning sequences reveals plastic rules for chromatin organization. *Genome research* 21, 1851-1862.
- Tsompana, M., and Buck, M.J. (2014). Chromatin accessibility: a window into the genome. *Epigenetics & chromatin* 7, 33.
- Valouev, A., Johnson, S.M., Boyd, S.D., Smith, C.L., Fire, A.Z., and Sidow, A. (2011). Determinants of nucleosome organization in primary human cells. *Nature* 474, 516-520.
- Vera, D.L., Madzima, T.F., Labonne, J.D., Alam, M.P., Hoffman, G.G., Girimurugan, S.B., Zhang, J., McGinnis, K.M., Dennis, J.H., and Bass, H.W. (2014). Differential nuclease sensitivity profiling of chromatin reveals biochemical footprints coupled to gene expression and functional DNA elements in maize. *The Plant cell* 26, 3883-3893.

- Vierstra, J., Wang, H., John, S., Sandstrom, R., and Stamatoyannopoulos, J.A. (2014). Coupling transcription factor occupancy to nucleosome architecture with DNase-FLASH. *Nature methods* 11, 66-72.
- Voong, L.N., Xi, L., Sebeson, A.C., Xiong, B., Wang, J.P., and Wang, X. (2016). Insights into Nucleosome Organization in Mouse Embryonic Stem Cells through Chemical Mapping. *Cell* 167, 1555-1570 e1515.
- Weber, C.M., Ramachandran, S., and Henikoff, S. (2014). Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. *Molecular cell* 53, 819-830.
- Weiner, A., Hsieh, T.H., Appleboim, A., Chen, H.V., Rahat, A., Amit, I., Rando, O.J., and Friedman, N. (2015). High-resolution chromatin dynamics during a yeast stress response. *Molecular cell* 58, 371-386.
- Weiner, A., Hughes, A., Yassour, M., Rando, O.J., and Friedman, N. (2010). High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome research* 20, 90-100.
- West, J.A., Cook, A., Alver, B.H., Stadtfeld, M., Deaton, A.M., Hochedlinger, K., Park, P.J., Tolstorukov, M.Y., and Kingston, R.E. (2014). Nucleosomal occupancy changes locally over key regulatory regions during cell differentiation and reprogramming. *Nature communications* 5, 4719.
- West, S.M., Rohs, R., Mann, R.S., and Honig, B. (2010). Electrostatic interactions between arginines and the minor groove in the nucleosome. *Journal of biomolecular structure & dynamics* 27, 861-866.
- Widlund, H.R., Vitolo, J.M., Thiriet, C., and Hayes, J.J. (2000). DNA sequence-dependent contributions of core histone tails to nucleosome stability: differential effects of acetylation and proteolytic tail removal. *Biochemistry* 39, 3835-3841.
- Wiechens, N., Singh, V., Gkikopoulos, T., Schofield, P., Rocha, S., and Owen-Hughes, T. (2016). The Chromatin Remodelling Enzymes SNF2H and SNF2L Position Nucleosomes adjacent to CTCF and Other Transcription Factors. *PLoS genetics* 12, e1005940.
- Workman, J.L., and Kingston, R.E. (1992). Nucleosome core displacement in vitro via a metastable transcription factor-nucleosome complex. *Science* 258, 1780-1784.
- Wu, C., Wong, Y.C., and Elgin, S.C. (1979). The chromatin structure of specific genes: II. Disruption of chromatin structure during gene activity. *Cell* 16, 807-814.
- Xi, Y., Yao, J., Chen, R., Li, W., and He, X. (2011). Nucleosome fragility reveals novel functional states of chromatin and poises genes for activation. *Genome research* 21, 718-724.
- Yin, H., Sweeney, S., Raha, D., Snyder, M., and Lin, H.F. (2011). A High-Resolution Whole-Genome Map of Key Chromatin Modifications in the Adult *Drosophila melanogaster*. *PLoS genetics* 7.
- Yuan, G.C., Liu, Y.J., Dion, M.F., Slack, M.D., Wu, L.F., Altschuler, S.J., and Rando, O.J. (2005). Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* 309, 626-630.
- Zaret, K.S., and Mango, S.E. (2016). Pioneer transcription factors, chromatin dynamics, and cell fate control. *Current opinion in genetics & development* 37, 76-81.
- Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., *et al.* (2008). Model-based analysis of ChIP-Seq (MACS). *Genome biology* 9, R137.
- Zhang, Y., Moqtaderi, Z., Rattner, B.P., Euskirchen, G., Snyder, M., Kadonaga, J.T., Liu, X.S., and Struhl, K. (2010). Evidence against a genomic code for nucleosome positioning. Reply to "Nucleosome sequence preferences influence in vivo nucleosome organization.". *Nature structural & molecular biology* 17, 920-923.

### *Bibliography*

Zhang, Z., and Pugh, B.F. (2011). High-resolution genome-wide mapping of the primary structure of chromatin. *Cell* **144**, 175-186.

Zhang, Z., Wippo, C.J., Wal, M., Ward, E., Korber, P., and Pugh, B.F. (2011). A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. *Science* **332**, 977-980.